CHAPTER 3

Graphical Methods for Describing Data

Florin Tirlea/iStockphoto

Most college students (and their parents) are concerned about the cost of a college education. *The Chronicle of Higher Education* (August 2008) reported the average tuition and fees for 4-year public institutions in each of the 50 U.S. states for the 2006-2007 academic year. Average tuition and fees (in dollars) are given for each state:

4712	4422	4669	4937	4452	4634	7151	7417	3050	3851
3930	4155	8038	6284	6019	4966	5821	3778	6557	7106
7629	7504	7392	4457	6320	5378	5181	2844	9003	9333
3943	5022	4038	5471	9010	4176	5598	9092	6698	7914
5077	5009	5114	3757	9783	6447	5636	4063	6048	2951

Make the most of your study time by accessing everything you need to succeed online with CourseMate.
Visit http://www.cengagebrain.com where you will find:

- An interactive eBook, which allows you to take notes, highlight, bookmark, search the text, and use in-context glossary definitions
- Step-by-step instructions for Minitab, Excel, TI-83/84, SPSS, and JMP
- Video solutions to selected exercises
- Data sets available for selected examples and exercises
- Online quizzes
- Flashcards
- Videos

Several questions could be posed about these data. What is a typical value of average tuition and fees for the 50 states? Are observations concentrated near the typical value, or does average tuition and fees differ quite a bit from state to state? Are there any states whose average tuition and fees are somehow unusual compared to the rest? What proportion of the states have average tuition and fees exceeding $6000? Exceeding $8000?

Questions such as these are most easily answered if the data can be organized in a sensible manner. In this chapter, we introduce some techniques for organizing and describing data using tables and graphs.

3.1 Displaying Categorical Data: Comparative Bar Charts and Pie Charts

Comparative Bar Charts

In Chapter 1 we saw that categorical data could be summarized in a frequency distribution and displayed graphically using a bar chart. Bar charts can also be used to give a visual comparison of two or more groups. This is accomplished by constructing two or more bar charts that use the same set of horizontal and vertical axes, as illustrated in Example 3.1.

EXAMPLE 3.1 How Far Is Far Enough

Each year The Princeton Review conducts a survey of high school students who are applying to college and parents of college applicants. The report "2009 College Hopes & Worries Survey Findings" (www.princetonreview.com/uploadedFiles/Test_Preparation/Hopes_and_Worries/colleg_hopes_worries_details.pdf) included a summary of how 12,715 high school students responded to the question "Ideally how far from home would you like the college you attend to be?" Also included was a summary of how 3007 parents of students applying to college responded to the question "How far from home would you like the college your child attends to be?" The accompanying relative frequency table summarized the student and parent responses.

	FREQUENCY		RELATIVE FREQUENCY	
Ideal Distance	Students	Parents	Students	Parents
Less than 250 miles	4450	1594	.35	.53
250 to 500 miles	3942	902	.31	.30
500 to 1000 miles	2416	331	.19	.11
More than 1000 miles	1907	180	.15	.06

When constructing a comparative bar chart we use the relative frequency rather than the frequency to construct the scale on the vertical axis so that we can make meaningful comparisons even if the sample sizes are not the same. The comparative bar chart for these data is shown in Figure 3.1. It is easy to see the differences between students and parents. A higher proportion of parents prefer a college close to home, and a higher

Step-by-Step technology instructions available online

proportion of students than parents believe that the ideal distance from home would be more than 500 miles.

To see why it is important to use relative frequencies rather than frequencies to compare groups of different sizes, consider the *incorrect* bar chart constructed using the frequencies rather than the relative frequencies (Figure 3.2). The incorrect bar chart conveys a very different and misleading impression of the differences between students and parents.

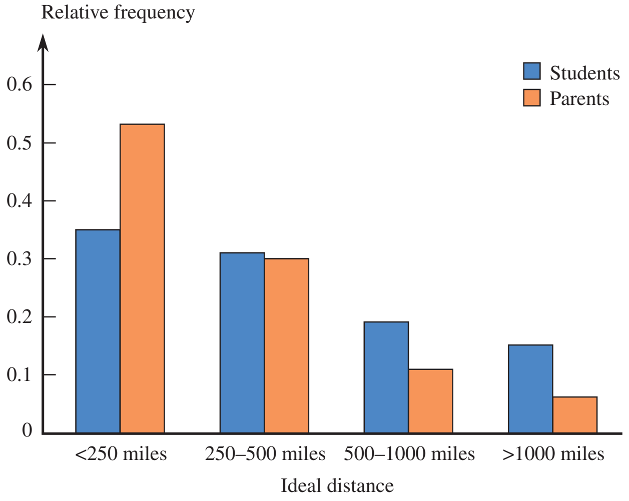


FIGURE 3.1
Comparative bar chart of ideal distance from home.

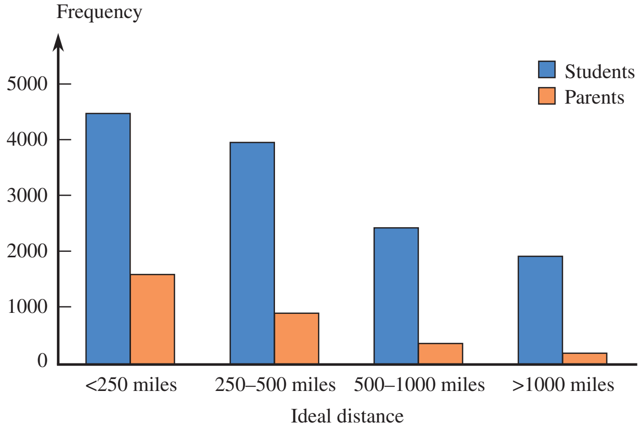


FIGURE 3.2
An *incorrect* comparative bar chart for the data of Example 3.1.

Pie Charts

A categorical data set can also be summarized using a pie chart. In a pie chart, a circle is used to represent the whole data set, with “slices” of the pie representing the possible categories. The size of the slice for a particular category is proportional to the corresponding frequency or relative frequency. Pie charts are most effective for summarizing data sets when there are not too many different categories.

EXAMPLE 3.2 Life Insurance for Cartoon Characters??

The article "Fred Flintstone, Check Your Policy" (*The Washington Post*, October 2, 2005) summarized the results of a survey of 1014 adults conducted by the Life and Health Insurance Foundation for Education. Each person surveyed was asked to select which of five fictional characters, Spider-Man, Batman, Fred Flintstone, Harry Potter, and Marge Simpson, he or she thought had the greatest need for life insurance. The resulting data are summarized in the pie chart of Figure 3.3.

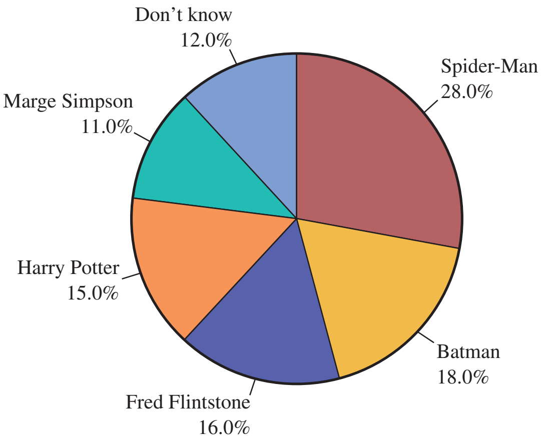


FIGURE 3.3
Pie chart of data on which fictional character most needs life insurance.

The survey results were quite different from an insurance expert's assessment. His opinion was that Fred Flintstone, a married father with a young child, was by far the one with the greatest need for life insurance. Spider-Man, unmarried with an elderly aunt, would need life insurance only if his aunt relied on him to supplement her income. Batman, a wealthy bachelor with no dependents, doesn't need life insurance in spite of his dangerous job!

Pie Chart for Categorical Data

When to Use Categorical data with a relatively small number of possible categories. Pie charts are most useful for illustrating proportions of the whole data set for various categories.

How to Construct

1. Draw a circle to represent the entire data set.
2. For each category, calculate the "slice" size. Because there are 360 degrees in a circle

   slice size $= 360 \cdot$ (category relative frequency)

3. Draw a slice of appropriate size for each category. This can be tricky, so most pie charts are generated using a graphing calculator or a statistical software package.

What to Look For

- Categories that form large and small proportions of the data set.

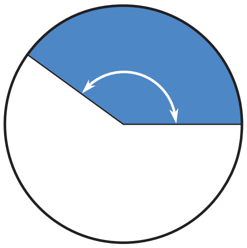
EXAMPLE 3.3 Watch Those Typos

Typos on a résumé do not make a very good impression when applying for a job. Senior executives were asked how many typos in a résumé would make them not consider a job candidate ("Job Seekers Need a Keen Eye," *USA Today*, August 3, 2009). The resulting data are summarized in the accompanying relative frequency distribution.

Number of Typos	Frequency	Relative Frequency
1	60	.40
2	54	.36
3	21	.14
4 or more	10	.07
Don't know	5	.03

To draw a pie chart by hand, we would first compute the slice size for each category. For the one typo category, the slice size would be

$$\text{slice size} = (.40)(360) = 144 \text{ degrees}$$


144 degrees, to represent first attempt category

We would then draw a circle and use a protractor to mark off a slice corresponding to about 144°, as illustrated here in the figure shown in the margin. Continuing to add slices in this way leads to a completed pie chart.

It is much easier to use a statistical software package to create pie charts than to construct them by hand. A pie chart for the typo data, created with the statistical software package Minitab, is shown in Figure 3.4.

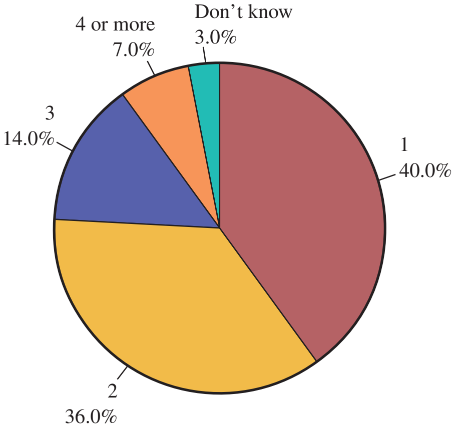


FIGURE 3.4
Pie chart for the typo data of Example 3.3.

Pie charts can be used effectively to summarize a single categorical data set if there are not too many different categories. However, pie charts are not usually the best tool if the goal is to compare groups on the basis of a categorical variable. This is illustrated in Example 3.4.

Step-by-Step technology instructions available online

EXAMPLE 3.4 Scientists and Nonscientists Do Not See Eye-to-Eye

Scientists and nonscientists were asked to indicate if they agreed or disagreed with the following statement: "When something is run by the government, it is usually inefficient and wasteful." The resulting data (from "Scientists, Public Differ in Outlooks," *USA Today*, July 10, 2009) were used to create the two pie charts in Figure 3.5.

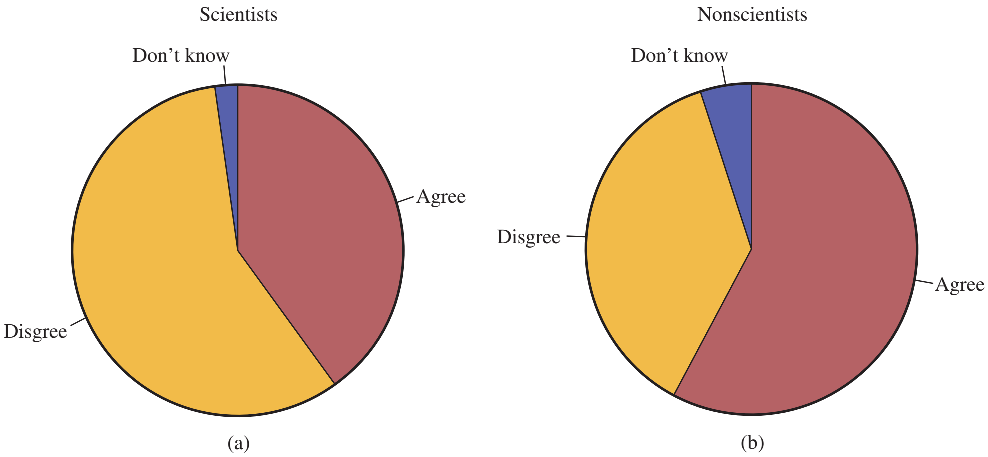


FIGURE 3.5
Pie charts for Example 3.4: (a) scientist data; (b) nonscientist data.

Although differences between scientists and nonscientists can be seen by comparing the pie charts of Figure 3.5, it can be difficult to compare category proportions using pie charts. A comparative bar chart (Figure 3.6) makes this type of comparison easier.

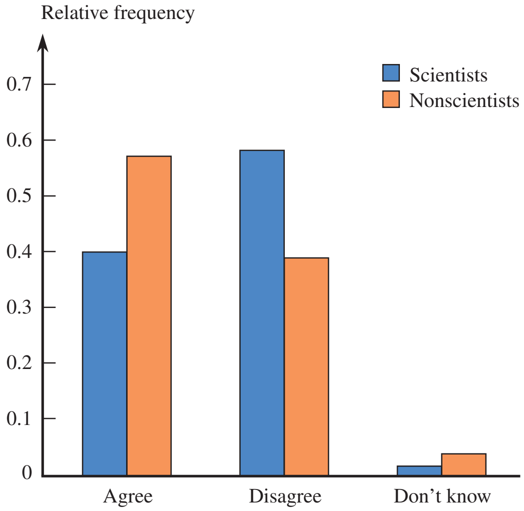


FIGURE 3.6
Comparative bar chart for the scientist and nonscientist data.

A Different Type of "Pie" Chart: Segmented Bar Graphs

A pie chart can be difficult to construct by hand, and the circular shape sometimes makes it difficult to compare areas for different categories, particularly when the relative frequencies for categories are similar. The **segmented bar graph** (also sometimes called a stacked bar graph) avoids these difficulties by using a rectangular bar rather than a circle to represent the entire data set. The bar is divided into segments, with different segments representing different categories. As with pie charts, the area of the segment for a particular category is proportional to the relative frequency for that category. Example 3.5 illustrates the construction of a segmented bar graph.

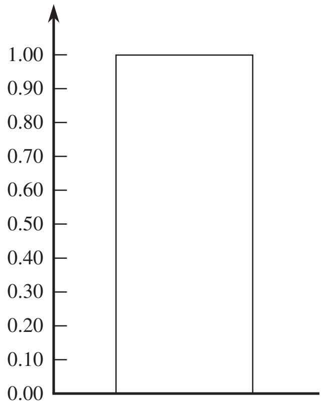
EXAMPLE 3.5 How College Seniors Spend Their Time

Each year, the Higher Education Research Institute conducts a survey of college seniors. In 2008, approximately 23,000 seniors participated in the survey ("Findings from the 2008 Administration of the College Senior Survey," Higher Education Research Institute, June 2009). The accompanying relative frequency table summarizes student response to the question: "During the past year, how much time did you spend studying and doing homework in a typical week?"

STUDYING/HOMEWORK

Amount of Time	Relative Frequency
2 hours or less	.074
3 to 5 hours	.227
6 to 10 hours	.285
11 to 15 hours	.181
16 to 20 hours	.122
Over 20 hours	.111

To construct a segmented bar graph for these data, first draw a bar of any fixed width and length, and then add a scale that ranges from 0 to 1, as shown.



Then divide the bar into six segments, corresponding to the six possible time categories in this example. The first segment, corresponding to the 2 hours or less category, ranges from 0 to .074. The second segment, corresponding to 3 to 5 hours, ranges from .074 to .301 (for a length of .227, the relative frequency for this category), and so on. The segmented bar graph is shown in Figure 3.7.

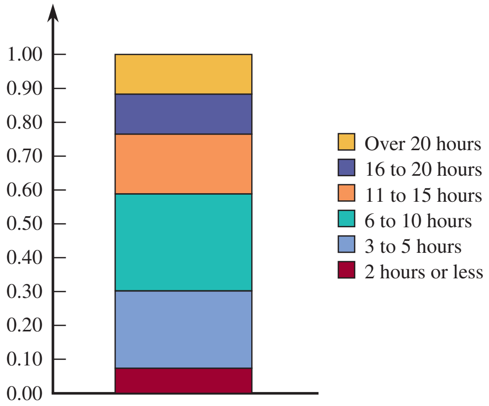


FIGURE 3.7
Segmented bar graph for the study time data of Example 3.5.

The same report also gave data on amount of time spent on exercise or sports in a typical week. Figure 3.8 shows horizontal segmented bar graphs (segmented bar graphs can be displayed either vertically or horizontally) for both time spent studying and time spent exercising. Viewing these graphs side by side makes it easy to see how students differ with respect to time spent on these two types of activities.

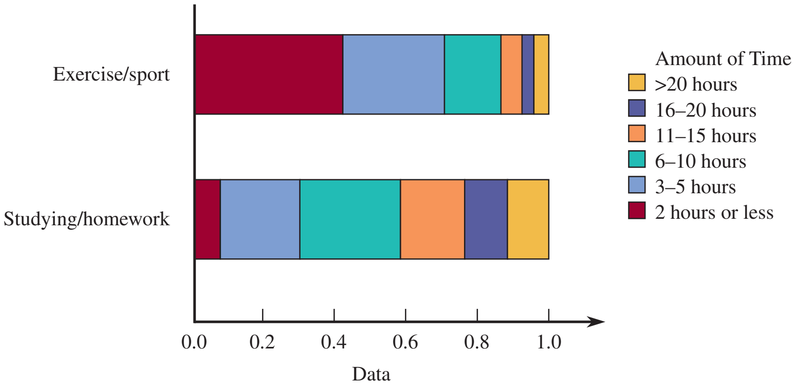


FIGURE 3.8
Segmented bar graphs for time spent studying and time spent exercising.

Other Uses of Bar Charts and Pie Charts

As we have seen in previous examples, bar charts and pie charts can be used to summarize categorical data sets. However, they are occasionally used for other purposes, as illustrated in Examples 3.6 and 3.7.

EXAMPLE 3.6 Grape Production

© PhotoLink/Photodisc/Getty Images

● The 2008 Grape Crush Report for California gave the following information on grape production for each of four different types of grapes (California Department of Food and Agriculture, March 10, 2009):

Type of Grape	Tons Produced
Red Wine Grapes	1,715,000
White Wine Grapes	1,346,000
Raisin Grapes	494,000
Table Grapes	117,000
Total	**3,672,000**

Although this table is not a frequency distribution, it is common to represent information of this type graphically using a pie chart, as shown in Figure 3.9. The pie represents the total grape production, and the slices show the proportion of the total production for each of the four types of grapes.

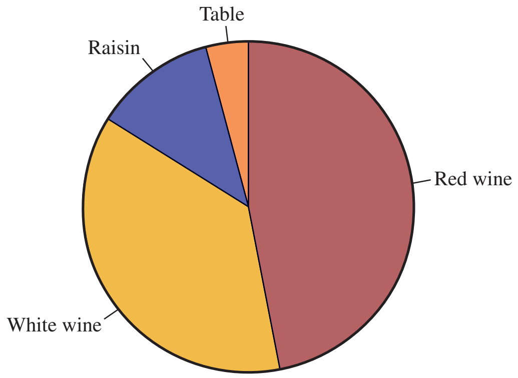


FIGURE 3.9
Pie chart for grape production data.

EXAMPLE 3.7 Back-to-College Spending

The National Retail Federation's 2008 Back to College Consumer Intentions and Actions Survey (www.nrf.com) asked each person in a sample of college students how much they planned to spend in various categories during the upcoming academic year. The average amounts of money (in dollars) that men and women planned to spend for five different types of purchases are shown in the accompanying table.

Type of Purchase	Average for Men	Average for Women
Clothing and Accessories	$207.46	$198.15
Shoes	$107.22	$88.65
School Supplies	$86.85	$81.56
Electronics and Computers	$533.17	$344.90
Dorm or Apartment Furnishings	$266.69	$266.98

● Data set available online

Even though this table is not a frequency distribution, this type of information is often represented graphically in the form of a bar chart, as illustrated in Figure 3.10. From the bar chart, we can see that the average amount of money that men and women plan to spend is similar for all of the types of purchases except for electronics and computers, in which the average for men is quite a bit higher than the average for women.

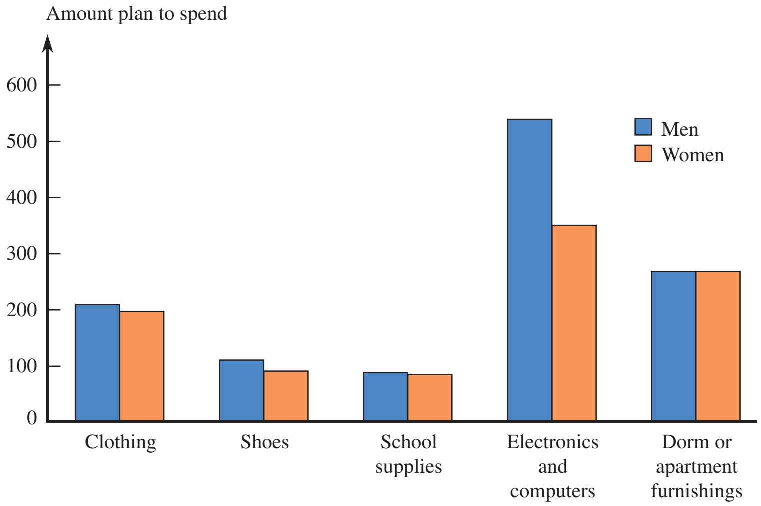


FIGURE 3.10
Comparative bar chart for the back-to-college spending data of men and women.

EXERCISES 3.1 - 3.14

3.1 Each person in a nationally representative sample of 1252 young adults age 23 to 28 years old was asked how they viewed their "financial physique" ("2009 Young Adults & Money Survey Findings," Charles Schwab, 2009). "Toned and fit" was chosen by 18% of the respondents, while 55% responded "a little bit flabby," and 27% responded "seriously out of shape." Summarize this information in a pie chart.

3.2 The accompanying graphical display appeared in *USA Today* (October 22, 2009). It summarizes survey responses to a question about whether visiting social networking sites is allowed at work. Which of the graph types introduced in this section is used to display the responses? (*USA Today* frequently adds artwork and text to their graphs to try to make them look more interesting.)

Image not available due to copyright restrictions

3.3 The survey referenced in the previous exercise was conducted by Robert Half Technology. This company issued a press release ("Whistle—But Don't Tweet—While You Work," www.roberthalftechnology.com, October 6, 2009) that provided more detail than in the *USA Today* snapshot graph. The actual question asked was “Which of the following most closely describes your company’s policy on visiting social networking sites, such as Facebook, MySpace and Twitter, while at work?” The responses are summarized in the following table:

Response Category	Relative Frequency (expressed as percent)
Prohibited completely	54%
Permitted for business purposes only	19%
Permitted for limited personal use	16%
Permitted for any type of personal use	10%
Don’t know/no answer	1%

a. Explain how the survey response categories and corresponding relative frequencies were used or modified to produce the graphical display in Exercise 3.2.

b. Using the original data in the table, construct a segmented bar graph.

c. What are two other types of graphical displays that would be appropriate for summarizing these data?

3.4 The National Confectioners Association asked 1006 adults the following question: “Do you set aside a personal stash of Halloween candy?” Fifty-five percent of those surveyed responded no, 41% responded yes, and 4% either did not answer the question or said they did not know (*USA Today*, October 22, 2009). Use the given information to construct a pie chart.

3.5 The report “Communicating to Teens (Aged 12–17)” (U.S. Department of Health and Human Services, www.cdc.gov) suggests that teens can be classified into five groups based on attitude, behavior, and conformity. The report also includes estimates of the percentage of teens who fall into each of these groups. The groups are described in the accompanying table.

Group and Description	Percentage of Teens in This Group
Explorer: creative, independent, and differs from the norm.	10%
Visible: well known and popular because of looks, personality or athletic ability	30%
Status Quo: display traditional values of moderation and achievement, seek mainstream acceptance	38%
Non-Teen: behave more like adults or young children because of lack of social skills or indifference to teen culture and style	14%
Isolator: psychologically isolated from both peers and adults	8%

Construct an appropriate graph to summarize the information in the table. Explain why you chose this particular type of graph.

3.6 The Center for Science in the Public Interest evaluated school cafeterias in 20 school districts across the United States. Each district was assigned a numerical score on the basis of rigor of food codes, frequency of food safety inspections, access to inspection information, and the results of cafeteria inspections. Based on the score assigned, each district was also assigned one of four grades. The scores and grades are summarized in the accompanying table, which appears in the report “Making the Grade: An Analysis of Food Safety in School Cafeterias” (cspi.us/new/pdf/makingthegrade.pdf, 2007).

🟦 Top of the Class 🟨 Passing 🟫 Barely Passing 🟧 Failing

Jurisdiction	Overall Score (out of 100)
City of Fort Worth, TX	80
King County, WA	79
City of Houston, TX	78
Maricopa County, AZ	77
City and County of Denver, CO	75
Dekalb County, GA	73
Farmington Valley Health District, CT	72
State of Virginia	72
Fulton County, GA	68
City of Dallas, TX	67
City of Philadelphia, PA	67
City of Chicago, IL	65
City and County of San Francisco, CA	64
Montgomery County, MD	63
Hillsborough County, FL	60
City of Minneapolis, MN	60
Dade County, FL	59
State of Rhode Island	54
District of Columbia	46
City of Hartford, CT	37

a. Two variables are summarized in the figure, grade and overall score. Is overall score a numerical or categorical variable? Is grade (indicated by the different colors in the figure) a numerical or categorical variable?

b. Explain how the figure is equivalent to a segmented bar graph of the grade data.

c. Construct a dotplot of the overall score data. Based on the dotplot, suggest an alternate assignment of grades (top of class, passing, etc.) to the 20 school districts. Explain the reasoning you used to make your assignment.

3.7 The article "Housework around the World" (*USA Today*, September 15, 2009) included the percentage of women who say their spouses never help with household chores for five different countries.

Country	Percentage
Japan	74%
France	44%
United Kingdom	40%
United States	34%
Canada	31%

a. Display the information in the accompanying table in a bar chart.

b. The article did not state how the author arrived at the given percentages. What are two questions that you would want to ask the author about how the data used to compute the percentages were collected?

c. Assuming that the data that were used to compute these percentages were collected in a reasonable way, write a few sentences describing how the five countries differ in terms of spouses helping their wives with housework.

3.8 The report "Findings from the 2008 Administration of the College Senior Survey" (Higher Education Research Institute, 2009) asked a large number of college seniors how they would rate themselves compared to the average person of their age with respect to physical health. The accompanying relative frequency table summarizes the responses for men and women.

Rating of Physical Health	Relative Frequency	
	Men	Women
Highest 10%	.220	.101
Above average	.399	.359
Average	.309	.449
Below average	.066	.086
Lowest 10%	.005	.005

a. Construct a comparative bar graph of the responses that allows you to compare the responses of men and women.

b. There were 8110 men and 15,260 women who responded to the survey. Explain why it is important that the comparative bar graph be constructed using the relative frequencies rather than the actual numbers of people (the frequencies) responding in each category.

c. Write a few sentences commenting on how college seniors perceive themselves with respect to physical health and how men and women differ in their perceptions.

3.9 The article "Rinse Out Your Mouth" (Associated Press, March 29, 2006) summarized results from a survey of 1001 adults on the use of profanity. When asked "How many times do you use swear words in conversations?" 46% responded a few or more times per week, 32% responded a few times a month or less, and 21% responded never. Use the given information to construct a segmented bar chart.

3.10 The article "The Need to Be Plugged In" (Associated Press, December 22, 2005) described the results of a survey of 1006 adults who were asked about various technologies, including personal computers, cell phones, and DVD players. The accompanying table summarizes the responses to questions about how essential these technologies were.

	Relative Frequency		
Response	Personal Computer	Cell Phone	DVD Player
Cannot imagine living without	.46	.41	.19
Would miss but could do without	.28	.25	.35
Could definitely live without	.26	.34	.46

Construct a comparative bar chart that shows the distribution of responses for the three different technologies.

3.11 ✦ Poor fitness in adolescents and adults increases the risk of cardiovascular disease. In a study of 3110 adolescents and 2205 adults (*Journal of the American Medical Association*, December 21, 2005), researchers found 33.6% of adolescents and 13.9% of adults were unfit; the percentage was similar in adolescent males (32.9%) and females (34.4%), but was higher in adult females (16.2%) than in adult males (11.8%).

a. Summarize this information using a comparative bar graph that shows differences between males and females within the two different age groups.

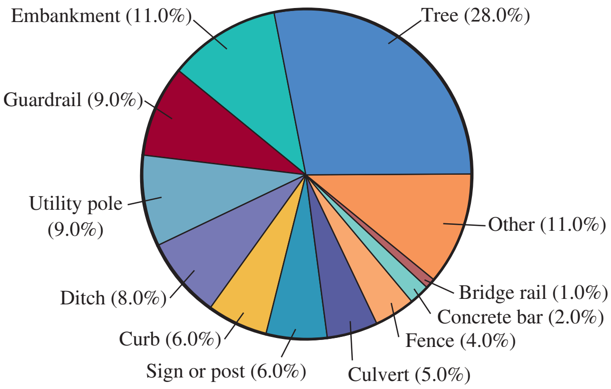
b. Comment on the interesting features of your graphical display.

3.12 A survey of 1001 adults taken by Associated Press–Ipsos asked "How accurate are the weather forecasts in your area?" (*San Luis Obispo Tribune*, June 15, 2005). The responses are summarized in the table below.

Extremely	4%
Very	27%
Somewhat	53%
Not too	11%
Not at all	4%
Not sure	1%

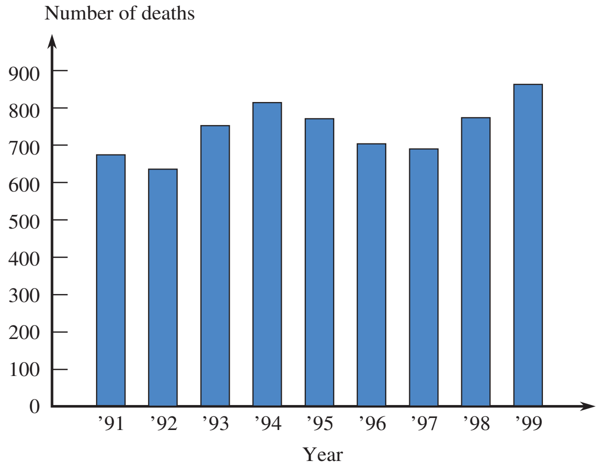
a. Construct a pie chart to summarize these data.
b. Construct a bar chart to summarize these data.
c. Which of these charts—a pie chart or a bar chart—best summarizes the important information? Explain.

3.13 In a discussion of accidental deaths involving roadside hazards, the web site highwaysafety.com included a pie chart like the one shown:



a. Do you think this is an effective use of a pie chart? Why or why not?
b. Construct a bar chart to show the distribution of deaths by object struck. Is this display more effective than the pie chart in summarizing this data set? Explain.

3.14 The article "Death in Roadwork Zones at Record High" (*San Luis Obispo Tribune*, July 25, 2001) included a bar chart similar to this one:



a. Comment on the trend over time in the number of people killed in highway work zones.
b. Would a pie chart have also been an effective way to summarize these data? Explain why or why not.

3.2 Displaying Numerical Data: Stem-and-Leaf Displays

A stem-and-leaf display is an effective and compact way to summarize univariate numerical data. Each number in the data set is broken into two pieces, a stem and a leaf. The **stem** is the first part of the number and consists of the beginning digit(s). The **leaf** is the last part of the number and consists of the final digit(s). For example, the number 213 might be split into a stem of 2 and a leaf of 13 or a stem of 21 and a leaf of 3. The resulting stems and leaves are then used to construct the display.

EXAMPLE 3.8 Should Doctors Get Auto Insurance Discounts?

● Many auto insurance companies give job-related discounts of between 5 and 15%. The article "Auto-Rate Discounts Seem to Defy Data" (*San Luis Obispo Tribune*, June 19, 2004) included the accompanying data on the number of automobile accidents per year for every 1000 people in 40 occupations.

Occupation	Accidents per 1000	Occupation	Accidents per 1000
Student	152	Banking-finance	89
Physician	109	Customer service	88
Lawyer	106	Manager	88
Architect	105	Medical support	87
Real estate broker	102	Computer-related	87
Enlisted military	199	Dentist	86
Social worker	198	Pharmacist	85
Manual laborer	196	Proprietor	84
Analyst	195	Teacher, professor	84
Engineer	194	Accountant	84
Consultant	194	Law enforcement	79
Sales	193	Physical therapist	78
Military officer	191	Veterinarian	78
Nurse	190	Clerical, secretary	77
School administrator	190	Clergy	76
Skilled labor	190	Homemaker	76
Librarian	190	Politician	76
Creative arts	190	Pilot	75
Executive	189	Firefighter	67
Insurance agent	189	Farmer	43

Figure 3.11 shows a stem-and-leaf display for the accident rate data.

The numbers in the vertical column on the left of the display are the **stems**. Each number to the right of the vertical line is a **leaf** corresponding to one of the observations in the data set. The legend

Stem: Tens
Leaf: Ones

tells us that the observation that had a stem of 4 and a leaf of 3 corresponds to an occupation with an accident rate of 43 per 1000 (as opposed to 4.3 or 0.43). Similarly, the observation with the stem of 10 and leaf of 2 corresponds to 102 accidents per 1000 (the leaf of 2 is the ones digit) and the observation with the stem of 15 and leaf of 2 corresponds to 152 accidents per 1000.

The display in Figure 3.11 suggests that a typical or representative value is in the stem 8 or 9 row, perhaps around 90. The observations are mostly concentrated in the 75 to 109 range, but there are a couple of values that stand out on the low end (43 and 67) and one observation (152) that is far removed from the rest of the data on the high end.

4	3
5	
6	7
7	56667889
8	44567788999
9	000013445689
10	2569
11	
12	
13	
14	
15	2

Stem: Tens
Leaf: Ones

FIGURE 3.11
Stem-and-leaf display for accident rate per 1000 for forty occupations

Step-by-step technology instructions available online

● Data set available online

From the point of view of an auto insurance company it might make sense to offer discounts to occupations with low accident rates—maybe farmers (43 auto accidents per 1000 farmers) or firefighters (67 accidents per 1000 firefighters) or even some of the occupations with accident rates in the 70s. The "discounts seem to defy data" in the title of the article refers to the fact that some insurers provide discounts to doctors and engineers, but not to homemakers, politicians, and other occupations with lower accident rates. Two possible explanations were offered for this apparent discrepancy. One is that it is possible that while some occupations have higher accident rates, they also have lower average cost per claim. Accident rates alone may not reflect the actual cost to the insurance company. Another possible explanation is that the insurance companies may offer the discounted auto insurance in order to attract people who would then also purchase other types of insurance such as malpractice or liability insurance.

The leaves on each line of the display in Figure 3.11 have been arranged in order from smallest to largest. Most statistical software packages order the leaves this way, but it is not necessary to do so to get an informative display that still shows many of the important characteristics of the data set, such as shape and spread.

Stem-and-leaf displays can be useful to get a sense of a typical value for the data set, as well as a sense of how spread out the values in the data set are. It is also easy to spot data values that are unusually far from the rest of the values in the data set. Such values are called outliers. The stem-and-leaf display of the accident rate data (Figure 3.11) shows an outlier on the low end (43) and an outlier on the high end (152).

DEFINITION

An **outlier** is an unusually small or large data value. A precise rule for deciding when an observation is an outlier is given in Chapter 4.

Stem-and-Leaf Displays

When to Use Numerical data sets with a small to moderate number of observations (does not work well for very large data sets)

How to Construct

1. Select one or more leading digits for the stem values. The trailing digits (or sometimes just the first one of the trailing digits) become the leaves.
2. List possible stem values in a vertical column.
3. Record the leaf for every observation beside the corresponding stem value.
4. Indicate the units for stems and leaves someplace in the display.

What to Look For The display conveys information about

- a representative or typical value in the data set
- the extent of spread about a typical value
- the presence of any gaps in the data
- the extent of symmetry in the distribution of values
- the number and location of peaks

EXAMPLE 3.9 Tuition at Public Universities

● The introduction to this chapter gave data on average tuition and fees at public institutions in the year 2007 for the 50 U.S. states. The observations ranged from a low value of 2844 to a high value of 9783. The data are reproduced here:

4712	4422	4669	4937	4452	4634	7151	7417	3050	3851
3930	4155	8038	6284	6019	4966	5821	3778	6557	7106
7629	7504	7392	4457	6320	5378	5181	2844	9003	9333
3943	5022	4038	5471	9010	4176	5598	9092	6698	7914
5077	5009	5114	3757	9783	6447	5636	4063	6048	2951

A natural choice for the stem is the leading (thousands) digit. This would result in a display with 7 stems (2, 3, 4, 5, 6, 7, 8, and 9). Using the first two digits of a number as the stem would result in 69 stems (28, 29, . . . , 97). A stem-and-leaf display with 56 stems would not be an effective summary of the data. *In general, stem-and-leaf displays that use between 5 and 20 stems tend to work well.*

If we choose the thousands digit as the stem, the remaining three digits (the hundreds, tens, and ones) would form the leaf. For example, for the first few values in the first column of data, we would have

4712 → stem = 4, leaf = 712
3930 → stem = 3, leaf = 930
7629 → stem = 7, leaf = 629

The leaves have been entered in the display of Figure 3.12 in the order they are encountered in the data set. Commas are used to separate the leaves only when each leaf has two or more digits. Figure 3.12 shows that most states had average tuition and fees in the $4000 to $7000 range and that the typical average tuition and fees is around $6000. A few states have average tuition and fees at public four-year institutions that are quite a bit higher than most other states (the five states with the highest values were Vermont, New Jersey, Pennsylvania, Ohio, and New Hampshire).

● Data set available online

2	844, 951
3	050, 851, 930, 778, 943, 757
4	712, 422, 669, 937, 452, 634, 155, 966, 457, 038, 176, 063
5	821, 378, 181, 022, 471, 598, 077, 009, 114, 636
6	284, 019, 557, 320, 698, 447, 048
7	151, 417, 106, 629, 504, 392, 914
8	038
9	003, 333, 010, 092, 783

Stem: Thousands
Leaf: Ones

FIGURE 3.12
Stem-and-leaf display of average tuition and fees.

An alternative display (Figure 3.13) results from dropping all but the first digit of the leaf. This is what most statistical computer packages do when generating a display; little information about typical value, spread, or shape is lost in this truncation and the display is simpler and more compact.

2	89
3	089797
4	746946194010
5	8310450016
6	2053640
7	1416539
8	0
9	03007

Stem: Thousands
Leaf: Hundreds

FIGURE 3.13
Stem-and-leaf display of the average tuition and fees data using truncated stems.

Repeated Stems to Stretch a Display

Sometimes a natural choice of stems gives a display in which too many observations are concentrated on just a few stems. A more informative picture can be obtained by dividing the leaves at any given stem into two groups: those that begin with 0, 1, 2, 3, or 4 (the "low" leaves) and those that begin with 5, 6, 7, 8, or 9 (the "high" leaves). Then each stem value is listed twice when constructing the display, once for the low leaves and once again for the high leaves. It is also possible to repeat a stem more than twice. For example, each stem might be repeated five times, once for each of the leaf groupings {0, 1}, {2, 3}, {4, 5}, {6, 7}, and {8, 9}.

EXAMPLE 3.10 Median Ages in 2030

● The accompanying data on the Census Bureau's projected median age in 2030 for the 50 U.S. states and Washington D.C. appeared in the article "2030 Forecast: Mostly Gray" (*USA Today*, April 21, 2005). The median age for a state is the age that divides the state's residents so that half are younger than the median age and half are older than the median age.

Projected Median Age

41.0	32.9	39.3	29.3	37.4	35.6	41.1	43.6	33.7	45.4	35.6	38.7
39.2	37.8	37.7	42.0	39.1	40.0	38.8	46.9	37.5	40.2	40.2	39.0
41.1	39.6	46.0	38.4	39.4	42.1	40.8	44.8	39.9	36.8	43.2	40.2
37.9	39.1	42.1	40.7	41.3	41.5	38.3	34.6	30.4	43.9	37.8	38.5
46.7	41.6	46.4									

The ages in the data set range from 29.3 to 46.9. Using the first two digits of each data value for the stem results in a large number of stems, while using only the first digit results in a stem-and-leaf display with only three stems.

The stem-and-leaf display using single digit stems and leaves truncated to a single digit is shown in Figure 3.14. A stem-and-leaf display that uses repeated stems is shown in Figure 3.15. Here each stem is listed twice, once for the low leaves (those beginning with 0, 1, 2, 3, 4) and once for the high leaves (those beginning with 5, 6, 7, 8, 9). This display is more informative than the one in Figure 3.14, but is much more compact than a display based on two-digit stems.

```
2 | 9
3 | 02345567777778888899999999
4 | 000000111111222333456666    Stem: Tens
                                Leaf:  Ones
```

FIGURE 3.14
Stem-and-leaf display for the projected median age data.

```
2H | 9
 3L | 0234
3H | 5567777778888899999999
 4L | 0000001111112223334
4H | 56666                      Stem: Tens
                                Leaf:  Ones
```

FIGURE 3.15
Stem-and-leaf display for the projected median age data using repeated stems.

● Data set available online

Comparative Stem-and-Leaf Displays

Frequently an analyst wishes to see whether two groups of data differ in some fundamental way. A comparative stem-and-leaf display, in which the leaves for one group are listed to the right of the stem values and the leaves for the second group are listed to the left, can provide preliminary visual impressions and insights.

EXAMPLE 3.11 Progress for Children

● The report "Progress for Children" (UNICEF, April 2005) included the accompanying data on the percentage of primary-school-age children who were enrolled in school for 19 countries in Northern Africa and for 23 countries in Central Africa.

Northern Africa

54.6 34.3 48.9 77.8 59.6 88.5 97.4 92.5 83.9 96.9 88.9
98.8 91.6 97.8 96.1 92.2 94.9 98.6 86.6

Central Africa

58.3 34.6 35.5 45.4 38.6 63.8 53.9 61.9 69.9 43.0 85.0
63.4 58.4 61.9 40.9 73.9 34.8 74.4 97.4 61.0 66.7 79.6
98.9

We will construct a comparative stem-and-leaf display using the first digit of each observation as the stem and the remaining two digits as the leaf. To keep the display simple the leaves will be truncated to one digit. For example, the observation 54.6 would be processed as

54.6 → stem = 5, leaf = 4 (truncated from 4.6)

and the observation 34.3 would be processed as

34.3 → stem = 3, leaf = 4 (truncated from 4.3)

The resulting comparative stem-and-leaf display is shown in Figure 3.16.

Central Africa		Northern Africa
4854	3	4
035	4	8
838	5	49
6113913	6	
943	7	76
5	8	8386
87	9	7268176248

Stem: Tens
Leaf: Ones

FIGURE 3.16
Comparative stem-and-leaf display for percentage of children enrolled in primary school.

From the comparative stem-and-leaf display you can see that there is quite a bit of variability in the percentage enrolled in school for both Northern and Central African countries and that the shapes of the two data distributions are quite different. The percentage enrolled in school tends to be higher in Northern African countries than in Central African countries, although the smallest value in each of the two data sets is about the same. For Northern African countries the distribution of values has a single peak in the 90s with the number of observations declining as we move toward the stems corresponding to lower percentages enrolled in school. For Central African countries the distribution is more symmetric, with a typical value in the mid 60s.

● Data set available online

EXERCISES 3.15 - 3.21

3.15 ● The U.S. Department of Health and Human Services provided the data in the accompanying table in the report "Births: Preliminary Data for 2007" (*National Vital Statistics Reports,* March 18, 2009). Entries in the table are the birth rates (births per 1,000 of population) for the year 2007.

State	Births per 1,000 of Population
Alabama	14.0
Alaska	16.2
Arizona	16.2
Arkansas	14.6
California	15.5
Colorado	14.6
Connecticut	11.9
Delaware	14.1
District of Columbia	15.1
Florida	13.1
Georgia	15.9
Hawaii	14.9
Idaho	16.7
Illinois	14.1
Indiana	14.2
Iowa	13.7
Kansas	15.1
Kentucky	14.0
Louisiana	15.4
Maine	10.7
Maryland	13.9
Massachusetts	12.1
Michigan	12.4
Minnesota	14.2
Mississippi	15.9
Missouri	13.9
Montana	13.0
Nebraska	15.2
Nevada	16.1
New Hampshire	10.8
New Jersey	13.4
New Mexico	15.5
New York	13.1
North Carolina	14.5
North Dakota	13.8
Ohio	13.2
Oklahoma	15.2
Oregon	13.2
Pennsylvania	12.1
Rhode Island	11.7
South Carolina	14.3
South Dakota	15.4
Tennessee	14.1
Texas	17.1
Utah	20.8
Vermont	10.5
Virginia	14.1
Washington	13.8
West Virginia	12.1
Wisconsin	13.0
Wyoming	15.1

(continued)

Construct a stem-and-leaf display using stems 10, 11 . . . 20. Comment on the interesting features of the display.

3.16 ● ✦ The National Survey on Drug Use and Health, conducted in 2006 and 2007 by the Office of Applied Studies, led to the following state estimates of the total number of people ages 12 and older who had used a tobacco product within the last month.

State	Number of People (in thousands)
Alabama	1,307
Alaska	161
Arizona	1,452
Arkansas	819
California	6,751
Colorado	1,171
Connecticut	766
Delaware	200
District of Columbia	141
Florida	4,392
Georgia	2,341
Hawaii	239
Idaho	305
Illinois	3,149
Indiana	1,740
Iowa	755
Kansas	726
Kentucky	1,294
Louisiana	1,138
Maine	347
Maryland	1,206
Massachusetts	1,427
Michigan	2,561
Minnesota	1,324

(continued)

State	Number of People (in thousands)
Mississippi	763
Missouri	1,627
Montana	246
Nebraska	429
Nevada	612
New Hampshire	301
New Jersey	1,870
New Mexico	452
New York	4,107
North Carolina	2,263
North Dakota	162
Ohio	3,256
Oklahoma	1,057
Oregon	857
Pennsylvania	3,170
Rhode Island	268
South Carolina	1,201
South Dakota	202
Tennessee	1,795
Texas	5,533
Utah	402
Vermont	158
Virginia	1,771
Washington	1,436
West Virginia	582
Wisconsin	1,504
Wyoming	157

a. Construct a stem-and-leaf display using thousands (of thousands) as the stems and truncating the leaves to the tens (of thousands) digit.

b. Write a few sentences describing the shape of the distribution and any unusual observations.

c. The four largest values were for California, Texas, Florida, and New York. Does this indicate that tobacco use is more of a problem in these states than elsewhere? Explain.

d. If you wanted to compare states on the basis of the extent of tobacco use, would you use the data in the given table? If yes, explain why this would be reasonable. If no, what would you use instead as the basis for the comparison?

3.17 ● The article "Going Wireless" (*AARP Bulletin*, June 2009) reported the estimated percentage of households with only wireless phone service (no land line) for the 50 U.S. states and the District of Columbia. In the accompanying data table, each state was also classified into one of three geographical regions—West (W), Middle states (M), and East (E).

Wireless %	Region	State
13.9	M	AL
11.7	W	AK
18.9	W	AZ
22.6	M	AR
9.0	W	CA
16.7	W	CO
5.6	E	CN
5.7	E	DE
20.0	E	DC
16.8	E	FL
16.5	E	GA
8.0	W	HI
22.1	W	ID
16.5	M	IL
13.8	M	IN
22.2	M	IA
16.8	M	KA
21.4	M	KY
15.0	M	LA
13.4	E	ME
10.8	E	MD
9.3	E	MA
16.3	M	MI
17.4	M	MN
19.1	M	MS
9.9	M	MO
9.2	W	MT
23.2	M	NE
10.8	W	NV
16.9	M	ND
11.6	E	NH
8.0	E	NJ
21.1	W	NM
11.4	E	NY
16.3	E	NC
14.0	E	OH
23.2	M	OK
17.7	W	OR
10.8	E	PA
7.9	E	RI
20.6	E	SC
6.4	M	SD
20.3	M	TN
20.9	M	TX
25.5	W	UT
10.8	E	VA
5.1	E	VT
16.3	W	WA
11.6	E	WV
15.2	M	WI
11.4	W	WY

a. Construct a stem-and-leaf display for the wireless percentage using the data from all 50 states and the District of Columbia. What is a typical value for this data set?

b. Construct a back-to-back stem-and-leaf display for the wireless percentage of the states in the West and the states in the East. How do the distributions of wireless percentages compare for states in the East and states in the West?

3.18 The article "Economy Low, Generosity High" (*USA Today*, July 28, 2009) noted that despite a weak economy in 2008, more Americans volunteered in their communities than in previous years. Based on census data (www.volunteeringinamerica.gov), the top and bottom five states in terms of percentage of the population who volunteered in 2008 were identified. The top five states were Utah (43.5%), Nebraska (38.9%), Minnesota (38.4%), Alaska (38.0%), and Iowa (37.1%). The bottom five states were New York (18.5%), Nevada (18.8%), Florida (19.6%), Louisiana (20.1%), and Mississippi (20.9%).

a. For the data set that includes the percentage who volunteered in 2008 for each of the 50 states, what is the largest value? What is the smallest value?

b. If you were going to construct a stem-and-leaf display for the data set consisting of the percentage who volunteered in 2008 for the 50 states, what stems would you use to construct the display? Explain your choice.

3.19 ● The article "Frost Belt Feels Labor Drain" (*USA Today*, May 1, 2008) points out that even though total population is increasing, the pool of young workers is shrinking in many states. This observation was prompted by the data in the accompanying table. Entries in the table are the percent change in the population of 25- to 44-year-olds over the period from 2000 to 2007. A negative percent change corresponds to a state that had fewer 25- to 44-year-olds in 2007 than in 2000 (a decrease in the pool of young workers).

State	% Change
Alabama	−4.1
Alaska	−2.5
Arizona	17.8
Arkansas	0.9
California	−0.4
Colorado	4.1
Connecticut	−9.9
Delaware	−2.2
District of Columbia	1.8
Florida	5.8
Georgia	7.2
Hawaii	−1.3
Idaho	11.1
Illinois	−4.6
Indiana	−3.1
Iowa	−6.5
Kansas	−5.3
Kentucky	−1.7
Louisiana	−11.9
Maine	−8.7
Maryland	−5.7
Massachusetts	−9.6
Michigan	−9.1
Minnesota	−4.5
Mississippi	−5.2
Missouri	−2.9
Montana	−3.7
Nebraska	−5.6
Nevada	22.0
New Hampshire	−7.5
New Jersey	−7.8
New Mexico	0.6
New York	−8.0
North Carolina	2.4
North Dakota	−10.9
Ohio	−8.2
Oklahoma	−1.6
Oregon	4.4
Pennsylvania	−9.1
Rhode Island	−8.8
South Carolina	0.1
South Dakota	−4.1
Tennessee	0.6
Texas	7.3
Utah	19.6
Vermont	−10.4
Virginia	−1.1
Washington	1.6
West Virginia	−5.4
Wisconsin	−5.0
Wyoming	−2.3

(continued)

a. The smallest value in the data set is −11.9 and the largest value is 22.0. One possible choice of stems for a stem-and-leaf display would be to use the tens digit, resulting in stems of −1, −0, 0, 1, and 2. Notice that because there are both negative and positive values in the data set, we would want to use two 0 stems—one where we can enter leaves for the negative percent changes that are between 0 and −9.9, and one where we could enter leaves for the positive percent changes that are between 0 and 9.9. Construct a stem-and-leaf plot using these five stems. (Hint: Think of each data value as having two digits before the decimal place, so 4.1 would be regarded as 04.1.)

b. Using two-digit stems would result in more than 30 stems, which is more than we would usually want for a stem-and-leaf display. Describe a strategy for using repeated stems that would result in a stem-and-leaf display with about 10 stems.

c. The article described "the frost belt" as the cold part of the country—the Northeast and Midwest—noting that states in the frost belt generally showed a decline in the number of people in the 25- to 44-year-old age group. How would you describe the group of states that saw a marked increase in the number of 25- to 44-year-olds?

3.20 ● ✦ A report from Texas Transportation Institute (Texas A&M University System, 2005) titled "Congestion Reduction Strategies" included the accompanying data on extra travel time for peak travel time in hours per year per traveler for different sized urban areas.

Very Large Urban Areas	Extra Hours per Year per Traveler
Los Angeles, CA	93
San Francisco, CA	72
Washington DC, VA, MD	69
Atlanta, GA	67
Houston, TX	63
Dallas, Fort Worth, TX	60
Chicago, IL-IN	58
Detroit, MI	57
Miami, FL	51
Boston, MA, NH, RI	51
New York, NY-NJ-CT	49
Phoenix, AZ	49
Philadelphia, PA-NJ-DE-MD	38

Large Urban Areas	Extra Hours per Year per Traveler
Riverside, CA	55
Orlando, FL	55
San Jose, CA	53
San Diego, CA	52
Denver, CO	51
Baltimore, MD	50
Seattle, WA	46
Tampa, FL	46
Minneapolis, St Paul, MN	43
Sacramento, CA	40
Portland, OR, WA	39
Indianapolis, IN	38
St Louis, MO-IL	35
San Antonio, TX	33
Providence, RI, MA	33
Las Vegas, NV	30
Cincinnati, OH-KY-IN	30
Columbus, OH	29
Virginia Beach, VA	26
Milwaukee, WI	23
New Orleans, LA	18
Kansas City, MO-KS	17
Pittsburgh, PA	14
Buffalo, NY	13
Oklahoma City, OK	12
Cleveland, OH	10

a. Construct a comparative stem-and-leaf plot for annual delay per traveler for each of the two different sizes of urban areas.

b. Is the following statement consistent with the display constructed in Part (a)? Explain.

> The larger the urban area, the greater the extra travel time during peak period travel.

3.21 ● High school dropout rates (percentages) for 2008 for the 50 states were given in the 2008 Kids Count Data Book (www.aecf.org) and are shown in the following table:

State	Rate
Alabama	8%
Alaska	10%
Arizona	9%
Arkansas	9%
California	6%
Colorado	8%
Connecticut	5%
Delaware	7%
Florida	7%
Georgia	8%
Hawaii	8%
Idaho	6%

(continued)

State	Rate
Illinois	6%
Indiana	8%
Iowa	3%
Kansas	5%
Kentucky	7%
Louisiana	10%
Maine	6%
Maryland	6%
Massachusetts	4%
Michigan	6%
Minnesota	3%
Mississippi	7%
Missouri	7%
Montana	9%
Nebraska	4%
Nevada	10%
New Hampshire	3%
New Jersey	4%
New Mexico	10%
New York	5%
North Carolina	8%
North Dakota	7%
Ohio	5%
Oklahoma	8%
Oregon	6%
Pennsylvania	5%
Rhode Island	6%
South Carolina	7%
South Dakota	6%
Tennessee	7%
Texas	7%
Utah	7%
Vermont	4%
Virginia	4%
Washington	7%
West Virginia	8%
Wisconsin	4%
Wyoming	6%

(continued)

Note that dropout rates range from a low of 3% to a high of 10%. In constructing a stem-and-leaf display for these data, if we regard each dropout rate as a two-digit number and use the first digit for the stem, then there are only two possible stems, 0 and 1. One solution is to use repeated stems. Consider a scheme that divides the leaf range into five parts: 0 and 1, 2 and 3, 4 and 5, 6 and 7, and 8 and 9. Then, for example, stem 0 could be repeated as

0 with leaves 0 and 1
0t with leaves 2 and 3
0f with leaves 4 and 5
0s with leaves 6 and 7
0* with leaves 8 and 9

Construct a stem-and-leaf display for this data set that uses stems 0t, 0f, 0s, 0*, and 1. Comment on the important features of the display.

3.3 Displaying Numerical Data: Frequency Distributions and Histograms

A stem-and-leaf display is not always an effective way to summarize data; it is unwieldy when the data set contains a large number of observations. Frequency distributions and histograms are displays that work well for large data sets.

Frequency Distributions and Histograms for Discrete Numerical Data

Discrete numerical data almost always result from counting. In such cases, each observation is a whole number. As in the case of categorical data, a frequency distribution for discrete numerical data lists each possible value (either individually or grouped into intervals), the associated frequency, and sometimes the corresponding relative frequency. Recall that relative frequency is calculated by dividing the frequency by the total number of observations in the data set.

EXAMPLE 3.12 Promiscuous Queen Bees

● Queen honey bees mate shortly after they become adults. During a mating flight, the queen usually takes multiple partners, collecting sperm that she will store and use throughout the rest of her life. The authors of the paper "The Curious Promiscuity of Queen Honey Bees" (*Annals of Zoology* [2001]: 255–265) studied the behavior of 30 queen honey bees to learn about the length of mating flights and the number of partners a queen takes during a mating flight. The accompanying data on number of partners were generated to be consistent with summary values and graphs given in the paper.

Number of Partners

12	2	4	6	6	7	8	7	8	11
8	3	5	6	7	10	1	9	7	6
9	7	5	4	7	4	6	7	8	10

The corresponding relative frequency distribution is given in Table 3.1. The smallest value in the data set is 1 and the largest is 12, so the possible values from 1 to 12 are listed in the table, along with the corresponding frequency and relative frequency.

TABLE 3.1 Relative Frequency Distribution for Number of Partners

Number of Partners	Frequency	Relative Frequency
1	1	.033
2	1	.033
3	1	.033
4	3	.100
5	2	.067
6	5	.167
7	7	.233
8	4	.133
9	2	.067
10	2	.067
11	1	.033
12	1	.033
Total	30	.999

$\frac{1}{30} = .033$

Differs from 1 due to rounding

From the relative frequency distribution, we can see that five of the queen bees had six partners during their mating flight. The corresponding relative frequency, $\frac{5}{30} = .167$, tells us that the proportion of queens with six partners is .167, or equivalently 16.7% of the queens had six partners. Adding the relative frequencies for the values 10, 11, and 12 gives

$$.067 + .033 + .033 = .133$$

indicating that 13.3% of the queens had 10 or more partners.

● Data set available online

It is possible to create a more compact frequency distribution by grouping some of the possible values into intervals. For example, we might group together 1, 2, and 3 partners to form an interval of 1–3, with a corresponding frequency of 3. The grouping of other values in a similar way results in the relative frequency distribution shown in Table 3.2.

TABLE 3.2 Relative Frequency Distribution of Number of Partners Using Intervals

Number of Partners	Frequency	Relative Frequency
1–3	3	.100
4–6	10	.333
7–9	13	.433
10–12	4	.133

A histogram for discrete numerical data is a graph of the frequency or relative frequency distribution, and it is similar to the bar chart for categorical data. Each frequency or relative frequency is represented by a rectangle centered over the corresponding value (or range of values) and the area of the rectangle is proportional to the corresponding frequency or relative frequency.

Histogram for Discrete Numerical Data

When to Use Discrete numerical data. Works well, even for large data sets.

How to Construct

1. Draw a horizontal scale, and mark the possible values of the variable.
2. Draw a vertical scale, and mark it with either frequency or relative frequency.
3. Above each possible value, draw a rectangle centered at that value (so that the rectangle for 1 is centered at 1, the rectangle for 5 is centered at 5, and so on). The height of each rectangle is determined by the corresponding frequency or relative frequency. Often possible values are consecutive whole numbers, in which case the base width for each rectangle is 1.

What to Look For

- Center or typical value
- Extent of spread or variability
- General shape
- Location and number of peaks
- Presence of gaps and outliers

EXAMPLE 3.13 Revisiting Promiscuous Queen Bees

The queen bee data of Example 3.12 were summarized in a frequency distribution. The corresponding histogram is shown in Figure 3.17. Note that each rectangle in the histogram is centered over the corresponding value. When relative frequency instead of frequency is used for the vertical scale, the scale on the vertical axis is different but all essential characteristics of the graph (shape, location, spread) are unchanged.

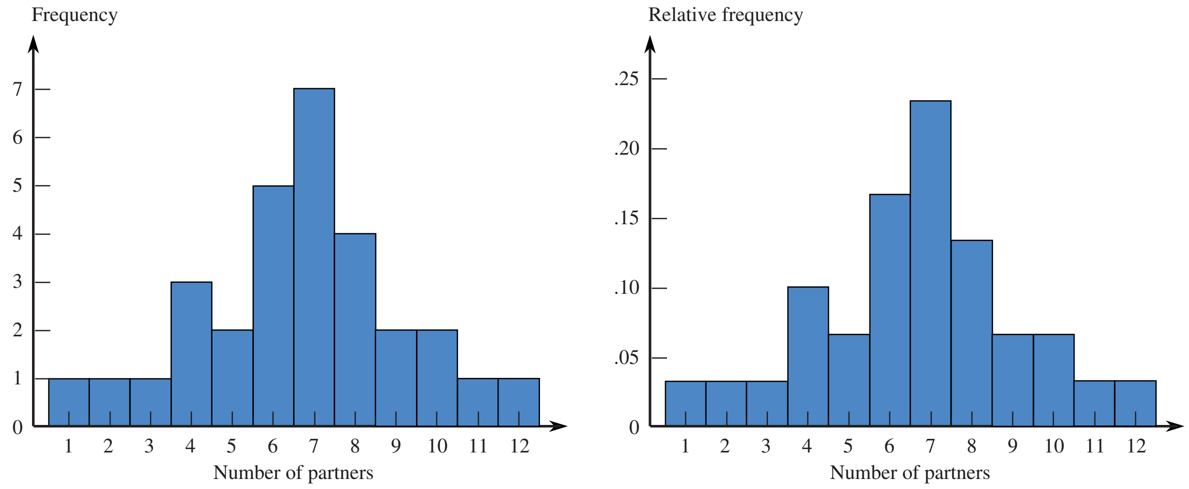


FIGURE 3.17
Histogram and relative frequency histogram of queen bee data.

A histogram based on the grouped frequency distribution of Table 3.2 can be constructed in a similar fashion, and is shown in Figure 3.18. A rectangle represents the frequency or relative frequency for each interval. For the interval 1–3, the rectangle extends from .5 to 3.5 so that there are no gaps between the rectangles of the histogram.

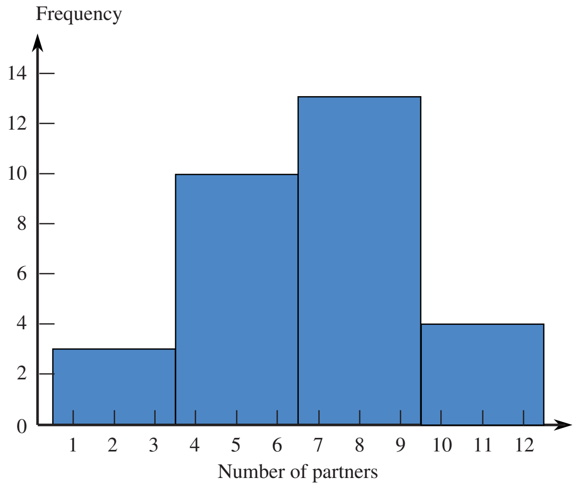


FIGURE 3.18
Histogram of queen bee data using intervals.

Sometimes a discrete numerical data set contains a large number of possible values and perhaps also has a few large or small values that are far away from most of the data. In this case, rather than forming a frequency distribution with a very long list of possible values, it is common to group the observed values into intervals or ranges. This is illustrated in Example 3.14.

EXAMPLE 3.14 Math SAT Score Distribution

Each of the 1,530,128 students who took the math portion of the SAT exam in 2009 received a score between 200 and 800. The score distribution was summarized in a frequency distribution table that appeared in the College Board report titled "2009 College Bound Seniors." A relative frequency distribution is given in Table 3.3 and

TABLE 3.3 Relative Frequency Distribution of Math SAT Score

Math SAT Score	Frequency	Relative Frequency
200–299	97,296	0.064
300–399	295,693	0.193
400–499	449,238	0.294
500–599	454,497	0.297
600–699	197,741	0.129
700–800	35,663	0.023

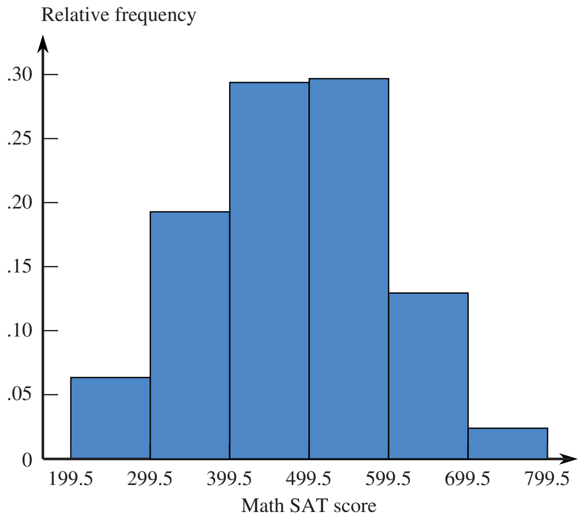


FIGURE 3.19 Relative frequency histogram for the math SAT data.

the corresponding relative frequency histogram is shown in Figure 3.19. Notice that rather than list each possible individual score value between 200 and 800, the scores are grouped into intervals (200 to 299, 300 to 399, etc.). This results in a much more compact table that still communicates the important features of the data set. Also, notice that because the data set is so large, the frequencies are also large numbers. Because of these large frequencies, it is easier to focus on the relative frequencies in our interpretation. From the relative frequency distribution and histogram, we can see that while there is a lot of variability in individual math SAT scores, the majority were in the 400 to 600 range and a typical value for math SAT looks to be something in the low 500s.

Before leaving this example, take a second look at the relative frequency histogram of Figure 3.19. Notice that there is one rectangle for each score interval in the relative frequency distribution. For simplicity we have chosen to treat the very last interval, 700 to 800, as if it were 700 to 799 so that all of the score ranges in the frequency distribution are the same width. Also note that the rectangle representing the score range 400 to 499 actually extends from 399.5 to 499.5 on the score scale. This is similar to what happens in histograms for discrete numerical data where there is no grouping. For example, in Figure 3.17 the rectangle representing 2 is centered at 2 but extends from 1.5 to 2.5 on the number of partners scale.

Frequency Distributions and Histograms for Continuous Numerical Data

The difficulty in constructing tabular or graphical displays with continuous data, such as observations on reaction time (in seconds) or weight of airline passenger carry-on luggage (in pounds), is that there are no natural categories. The way out of this dilemma is to define our own categories. For carry-on luggage weight, we might expect weights up to about 30 pounds. One way to group the weights into 5-pound intervals is shown in Figure 3.20. Then each observed data value could be classified into one of these intervals. The intervals used are sometimes called **class intervals**. The class intervals play the same role that the categories or individual values played in frequency distributions for categorical or discrete numerical data.

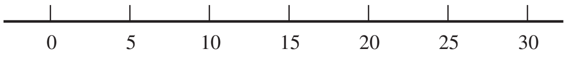


FIGURE 3.20
Suitable class intervals for carry-on luggage weight data.

There is one further difficulty we need to address. Where should we place an observation such as 20, which falls on a boundary between classes? Our convention is to define intervals so that such an observation is placed in the upper rather than the lower class interval. Thus, in a frequency distribution, one class might be 15 to <20, where the symbol < is a substitute for the phrase *less than*. This class will contain all observations that are greater than or equal to 15 and less than 20. The observation 20 would then fall in the class 20 to <25.

EXAMPLE 3.15 Enrollments at Public Universities

● States differ widely in the percentage of college students who are enrolled in public institutions. The National Center for Education Statistics provided the accompanying data on this percentage for the 50 U.S. states for fall 2007.

Percentage of College Students Enrolled in Public Institutions

96	86	81	84	77	90	73	53	90	96	73
93	76	86	78	76	88	86	87	64	60	58
89	86	80	66	70	90	89	82	73	81	73
72	56	55	75	77	82	83	79	75	59	59
43	50	64	80	82	75					

The smallest observation is 46 (Massachusetts) and the largest is 96 (Alaska and Wyoming). It is reasonable to start the first class interval at 40 and let each interval have a width of 10. This gives class intervals of 40 to <50, 50 to <60, 60 to <70, 70 to <80, 80 to <90, and 90 to <100.

Table 3.4 displays the resulting frequency distribution, along with the relative frequencies.

● Data set available online

TABLE 3.4 Frequency Distribution for Percentage of College Students Enrolled in Public Institutions

Class Interval	Frequency	Relative Frequency
40 to <50	1	.02
50 to <60	7	.14
60 to <70	4	.08
70 to <80	15	.30
80 to <90	17	.34
90 to <100	6	.12
	50	1.00

Various relative frequencies can be combined to yield other interesting information. For example,

$$\left(\begin{array}{l}\text{proportion of states}\\ \text{with percent in public}\\ \text{institutions less than 60}\end{array}\right) = \left(\begin{array}{l}\text{proportion in 40}\\ \text{to }{<}50\text{ class}\end{array}\right) + \left(\begin{array}{l}\text{proportion in 50}\\ \text{to }{<}60\text{ class}\end{array}\right)$$

$$= .02 + .14 = .16 \quad (16\%)$$

and

$$\left(\begin{array}{l}\text{proportion of states}\\ \text{with percent in}\\ \text{public institutions}\\ \text{between 60 and 90}\end{array}\right) = \left(\begin{array}{l}\text{proportion}\\ \text{in 60 to}\\ {<}70\text{ class}\end{array}\right) + \left(\begin{array}{l}\text{proportion}\\ \text{in 70 to}\\ {<}80\text{ class}\end{array}\right) + \left(\begin{array}{l}\text{proportion}\\ \text{in 80 to}\\ {<}90\text{ class}\end{array}\right)$$

$$= .08 + .30 + .34 = .72 \quad (72\%)$$

There are no set rules for selecting either the number of class intervals or the length of the intervals. Using a few relatively wide intervals will bunch the data, whereas using a great many relatively narrow intervals may spread the data over too many intervals, so that no interval contains more than a few observations. Neither type of distribution will give an informative picture of how values are distributed over the range of measurement, and interesting features of the data set may be missed. In general, with a small amount of data, relatively few intervals, perhaps between 5 and 10, should be used. With a large amount of data, a distribution based on 15 to 20 (or even more) intervals is often recommended. The quantity

$$\sqrt{\text{number of observations}}$$

is often used as an estimate of an appropriate number of intervals: 5 intervals for 25 observations, 10 intervals when the number of observations is 100, and so on.

Two people making reasonable and similar choices for the number of intervals, their width, and the starting point of the first interval will usually obtain similar histograms of the data.

Histograms for Continuous Numerical Data

When the class intervals in a frequency distribution are all of equal width, it is easy to construct a histogram using the information in a frequency distribution.

Histogram for Continuous Numerical Data When the Class Interval Widths are Equal

When to Use Continuous numerical data. Works well, even for large data sets.

How to Construct

1. Mark the boundaries of the class intervals on a horizontal axis.
2. Use either frequency or relative frequency on the vertical axis.
3. Draw a rectangle for each class directly above the corresponding interval (so that the edges are at the class interval boundaries). The height of each rectangle is the frequency or relative frequency of the corresponding class interval.

What to Look For

- Center or typical value
- Extent of spread, variability
- General shape
- Location and number of peaks
- Presence of gaps and outliers

EXAMPLE 3.16 TV Viewing Habits of Children

The article "Early Television Exposure and Subsequent Attention Problems in Children" (*Pediatrics,* April 2004) investigated the television viewing habits of children in the United States. Table 3.5 gives approximate relative frequencies (read from graphs that appeared in the article) for the number of hours spent watching TV per day for a sample of children at age 1 year and a sample of children at age 3 years. The data summarized in the article were obtained as part of a large scale national survey.

TABLE 3.5 Relative Frequency Distribution for Number of Hours Spent Watching TV per Day

TV Hours per Day	Age 1 Year Relative Frequency	Age 3 Years Relative Frequency
0 to <2	.270	.630
2 to <4	.390	.195
4 to <6	.190	.100
6 to <8	.085	.025
8 to <10	.030	.020
10 to <12	.020	.015
12 to <14	.010	.010
14 to <16	.005	.005

Figure 3.21(a) is the relative frequency histogram for the 1-year-old children and Figure 3.21(b) is the relative frequency histogram for 3-year-old children. Notice that both histograms have a single peak with the majority of children in both age groups concentrated in the smaller TV hours intervals. Both histograms are quite stretched out at the upper end, indicating some young children watch a lot of TV.

The big difference between the two histograms is at the low end, with a much higher proportion of 3-year-old children falling in the 0 to 2 TV hours interval than is the case for 1-year-old children. A typical number of TV hours per day for 1-year-old children would be somewhere between 2 and 4 hours, whereas a typical number of TV hours for 3-year-old children is in the 0 to 2 hours interval.

Step-by-step technology instructions available online

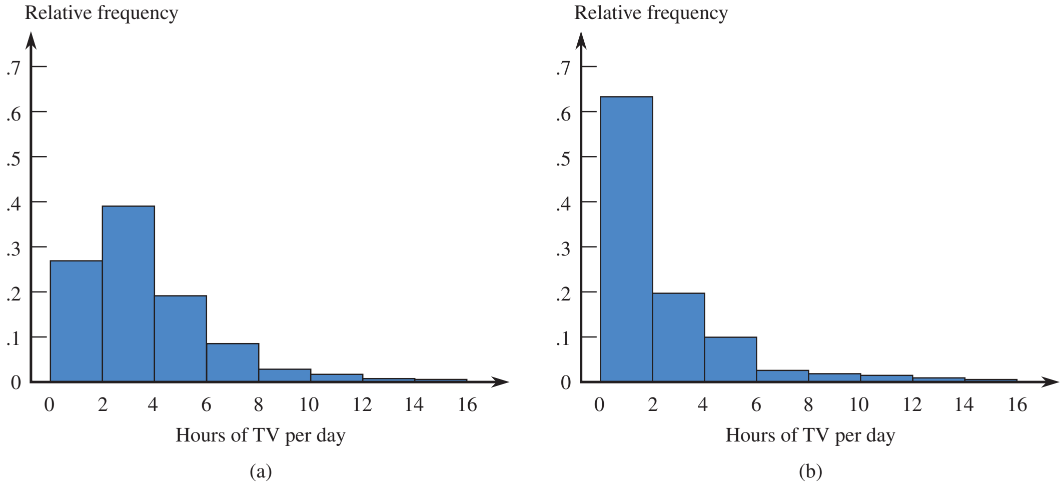


FIGURE 3.21
Histogram of TV hours per day: (a) 1-year-old children; (b) 3-year-old children.

Class Intervals of Unequal Widths Figure 3.22 shows a data set in which a great many observations are concentrated at the center of the data set, with only a few unusual, or stray, values both below and above the main body of data. If a frequency distribution is based on short intervals of equal width, a great many intervals will be required to capture all observations, and many of them will contain no observations, as shown in Figure 3.22(a). On the other hand, only a few wide intervals will capture all values, but then most of the observations will be grouped into a few intervals, as shown in Figure 3.22(b). In such situations, it is best to use a combination of wide class intervals where there are few data points and shorter intervals where there are many data points, as shown in Figure 3.22(c).

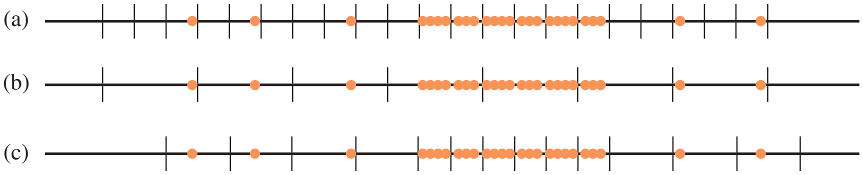


FIGURE 3.22
Three choices of class intervals for a data set with outliers: (a) many short intervals of equal width; (b) a few wide intervals of equal width; (c) intervals of unequal width.

Constructing a Histogram for Continuous Data When Class Interval Widths are Unequal

When class intervals are not of equal width, frequencies or relative frequencies should not be used on the vertical axis. Instead, the height of each rectangle, called the **density** for the class interval, is given by

$$\text{density} = \text{rectangle height} = \frac{\text{relative frequency of class interval}}{\text{class interval width}}$$

The vertical axis is called the **density scale**.

The use of the density scale to construct the histogram ensures that the area of each rectangle in the histogram will be proportional to the corresponding relative frequency. The formula for density can also be used when class widths are equal. However, when the intervals are of equal width, the extra arithmetic required to obtain the densities is unnecessary.

EXAMPLE 3.17 Misreporting Grade Point Average

When people are asked for the values of characteristics such as age or weight, they sometimes shade the truth in their responses. The article "Self-Reports of Academic Performance" (*Social Methods and Research* [November 1981]: 165–185) focused on such characteristics as SAT scores and grade point average (GPA). For each student in a sample, the difference in GPA (reported – actual) was determined. Positive differences resulted from individuals reporting GPAs larger than the correct values. Most differences were close to 0, but there were some rather large errors. Because of this, the frequency distribution based on unequal class widths shown in Table 3.6 gives an informative yet concise summary.

TABLE 3.6 Frequency Distribution for Errors in Reported GPA

Class Interval	Relative Frequency	Width	Density
-2.0 to <-0.4	.023	1.6	0.014
-0.4 to <-0.2	.055	.2	0.275
-0.2 to <-0.1	.097	.1	0.970
-0.1 to <0	.210	.1	2.100
0 to <0.1	.189	.1	1.890
0.1 to <0.2	.139	.1	1.390
0.2 to <0.4	.116	.2	0.580
0.4 to <2.0	.171	1.6	0.107

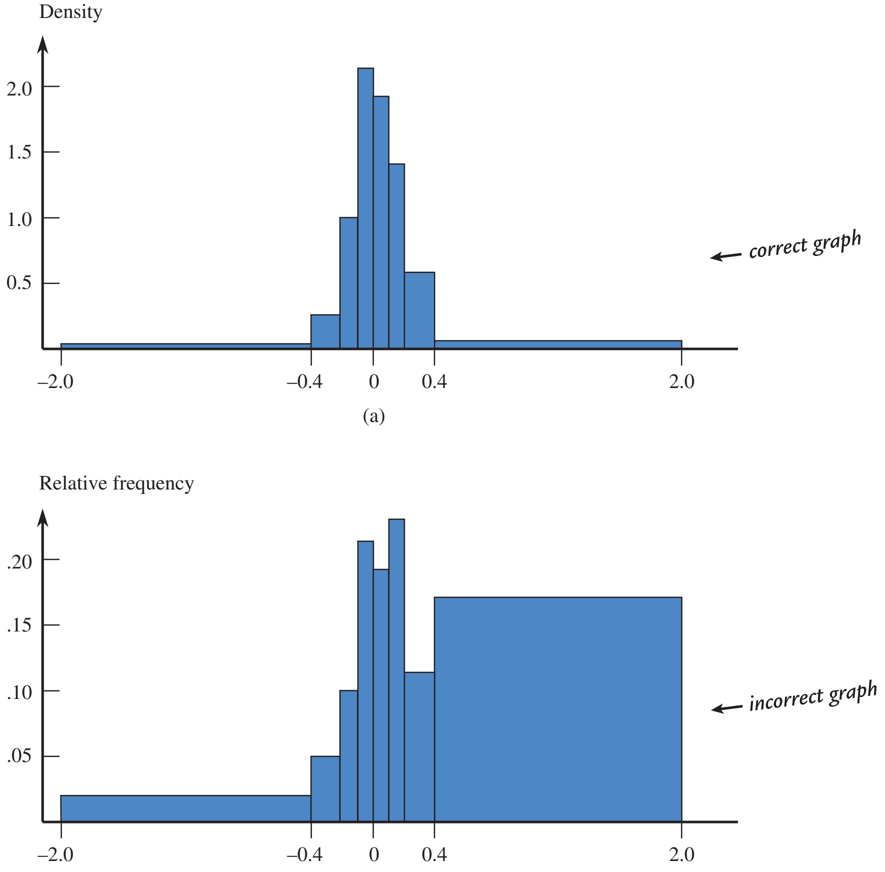


FIGURE 3.23 Histograms for errors in reporting GPA: (a) a correct histogram (height = density); (b) an incorrect histogram (height = relative frequency).

Figure 3.23 displays two histograms based on this frequency distribution. The histogram in Figure 3.23(a) is correctly drawn, with density used to determine the height of each bar. The histogram in Figure 3.23(b) has height equal to relative frequency and is therefore not correct. In particular, this second histogram considerably exaggerates the incidence of grossly overreported and underreported values—the areas of the two most extreme rectangles are much too large. The eye is naturally drawn to large areas, so it is important that the areas correctly represent the relative frequencies.

Histogram Shapes

General shape is an important characteristic of a histogram. In describing various shapes it is convenient to approximate the histogram itself with a smooth curve (called a *smoothed histogram*). This is illustrated in Figure 3.24.

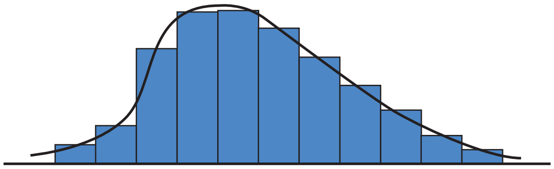


FIGURE 3.24
Approximating a histogram with a smooth curve.

One description of general shape relates to the number of peaks, or **modes**.

DEFINITION

A histogram is said to be **unimodal** if it has a single peak, **bimodal** if it has two peaks, and **multimodal** if it has more than two peaks.

These shapes are illustrated in Figure 3.25.

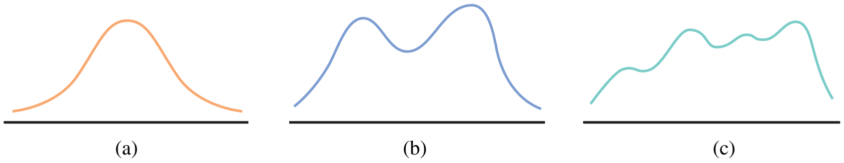


FIGURE 3.25
Smoothed histograms with various numbers of modes: (a) unimodal; (b) bimodal; (c) multimodal.

Bimodality sometimes occurs when the data set consists of observations on two quite different kinds of individuals or objects. For example, consider a large data set consisting of driving times for automobiles traveling between San Luis Obispo, California, and Monterey, California. This histogram would show two peaks, one for those cars that took the inland route (roughly 2.5 hours) and another for those cars traveling up the coast highway (3.5–4 hours). However, bimodality does not automatically follow in such situations. Bimodality will occur in the histogram of the combined groups only if the centers of the two separate histograms are far apart relative to the variability in the two data sets. Thus, a large data set consisting of heights of college students would probably not produce a bimodal histogram because the typical height for males (about 69 in.) and the typical height for females (about 66 in.) are not very far apart. Many histograms encountered in practice are unimodal, and multimodality is not as common.

Unimodal histograms come in a variety of shapes. A unimodal histogram is **symmetric** if there is a vertical line of symmetry such that the part of the histogram to the left of the line is a mirror image of the part to the right. (Bimodal and multimodal

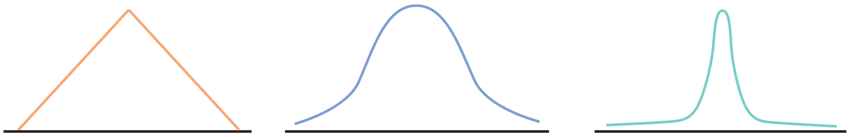


FIGURE 3.26
Several symmetric unimodal smoothed histograms.

histograms can also be symmetric in this way.) Several different symmetric smoothed histograms are shown in Figure 3.26.

Proceeding to the right from the peak of a unimodal histogram, we move into what is called the **upper tail** of the histogram. Going in the opposite direction moves us into the **lower tail**.

DEFINITION

A unimodal histogram that is not symmetric is said to be **skewed**. If the upper tail of the histogram stretches out much farther than the lower tail, then the distribution of values is **positively skewed** or **right skewed**. If, on the other hand, the lower tail is much longer than the upper tail, the histogram is **negatively skewed** or **left skewed**.

These two types of skewness are illustrated in Figure 3.27. Positive skewness is much more frequently encountered than is negative skewness. An example of positive skewness occurs in the distribution of single-family home prices in Los Angeles County; most homes are moderately priced (at least for California), whereas the relatively few homes in Beverly Hills and Malibu have much higher price tags.

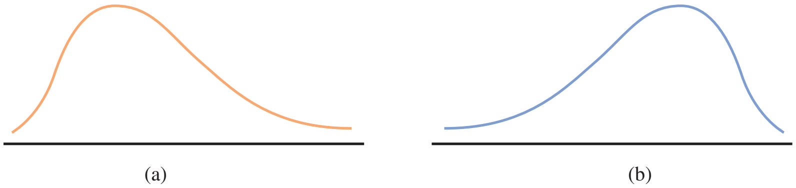


FIGURE 3.27
Two examples of skewed smoothed histograms: (a) positive skew; (b) negative skew.

One rather specific shape, a **normal curve**, arises more frequently than any other in statistical applications. Many histograms can be well approximated by a normal curve (for example, characteristics such as arm span and the weight of an apple). Here we briefly mention several of the most important qualitative properties of normal curves, postponing a more detailed discussion until Chapter 7. A normal curve is both symmetric and bell-shaped; it looks like the curve in Figure 3.28(a). However, not all bell-shaped curves are normal. In a normal curve, starting from the top of the bell the height of the curve decreases at a well-defined rate when moving toward either tail. (This rate of decrease is specified by a certain mathematical function.)

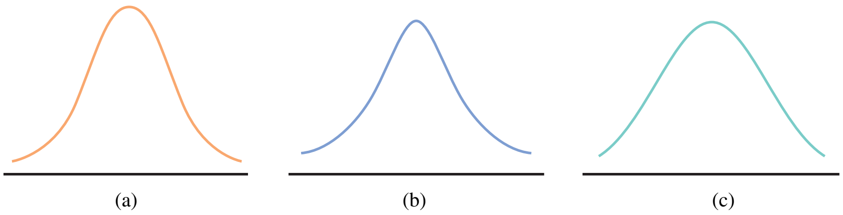


FIGURE 3.28
Three examples of bell-shaped histograms: (a) normal; (b) heavy-tailed; (c) light-tailed.

A curve with tails that do not decline as rapidly as the tails of a normal curve is called **heavy-tailed** (compared to the normal curve). Similarly, a curve with tails that decrease more rapidly than the normal tails is called **light-tailed**. Figures 3.28(b) and 3.28(c) illustrate these possibilities. The reason that we are concerned about the tails in a distribution is that many inferential procedures that work well (i.e., they result in accurate conclusions) when the population distribution is approximately normal perform poorly when the population distribution is heavy-tailed.

Do Sample Histograms Resemble Population Histograms?

Sample data are usually collected to make inferences about a population. The resulting conclusions may be in error if the sample is unrepresentative of the population. So how similar might a histogram of sample data be to the histogram of all population values? Will the two histograms be centered at roughly the same place and spread out to about the same extent? Will they have the same number of peaks, and will the peaks occur at approximately the same places?

A related issue concerns the extent to which histograms based on different samples from the same population resemble one another. If two different sample histograms can be expected to differ from one another in obvious ways, then at least one of them might differ substantially from the population histogram. If the sample differs substantially from the population, conclusions about the population based on the sample are likely to be incorrect. **Sampling variability**—the extent to which samples differ from one another and from the population—is a central idea in statistics. Example 3.18 illustrates sampling variability in histogram shapes.

EXAMPLE 3.18 What You Should Know About Bus Drivers . . .

● A sample of 708 bus drivers employed by public corporations was selected, and the number of traffic accidents in which each bus driver was involved during a 4-year period was determined ("Application of Discrete Distribution Theory to the Study of Noncommunicable Events in Medical Epidemiology," in *Random Counts in Biomedical and Social Sciences*, G. P. Patil, ed. [University Park, PA: Pennsylvania State University Press, 1970]). A listing of the 708 sample observations might look like this:

3 0 6 0 0 2 1 4 1 . . . 6 0 2

The frequency distribution (Table 3.7) shows that 117 of the 708 drivers had no accidents, a relative frequency of 117/708 = .165 (or 16.5%). Similarly, the proportion

TABLE 3.7 Frequency Distribution for Number of Accidents by Bus Drivers

Number of Accidents	Frequency	Relative Frequency
0	117	.165
1	157	.222
2	158	.223
3	115	.162
4	78	.110
5	44	.062
6	21	.030
7	7	.010
8	6	.008
9	1	.001
10	3	.004
11	1	.001
	708	.998

● Data set available online

of sampled drivers who had 1 accident is .222 (or 22.2%). The largest sample observation was 11.

Although the 708 observations actually constituted a sample from the population of all bus drivers, we will regard the 708 observations as constituting the entire population. The first histogram in Figure 3.29, then, represents the population histogram. The other four histograms in Figure 3.29 are based on four different samples of 50 observations each selected at random from this population. The five histograms

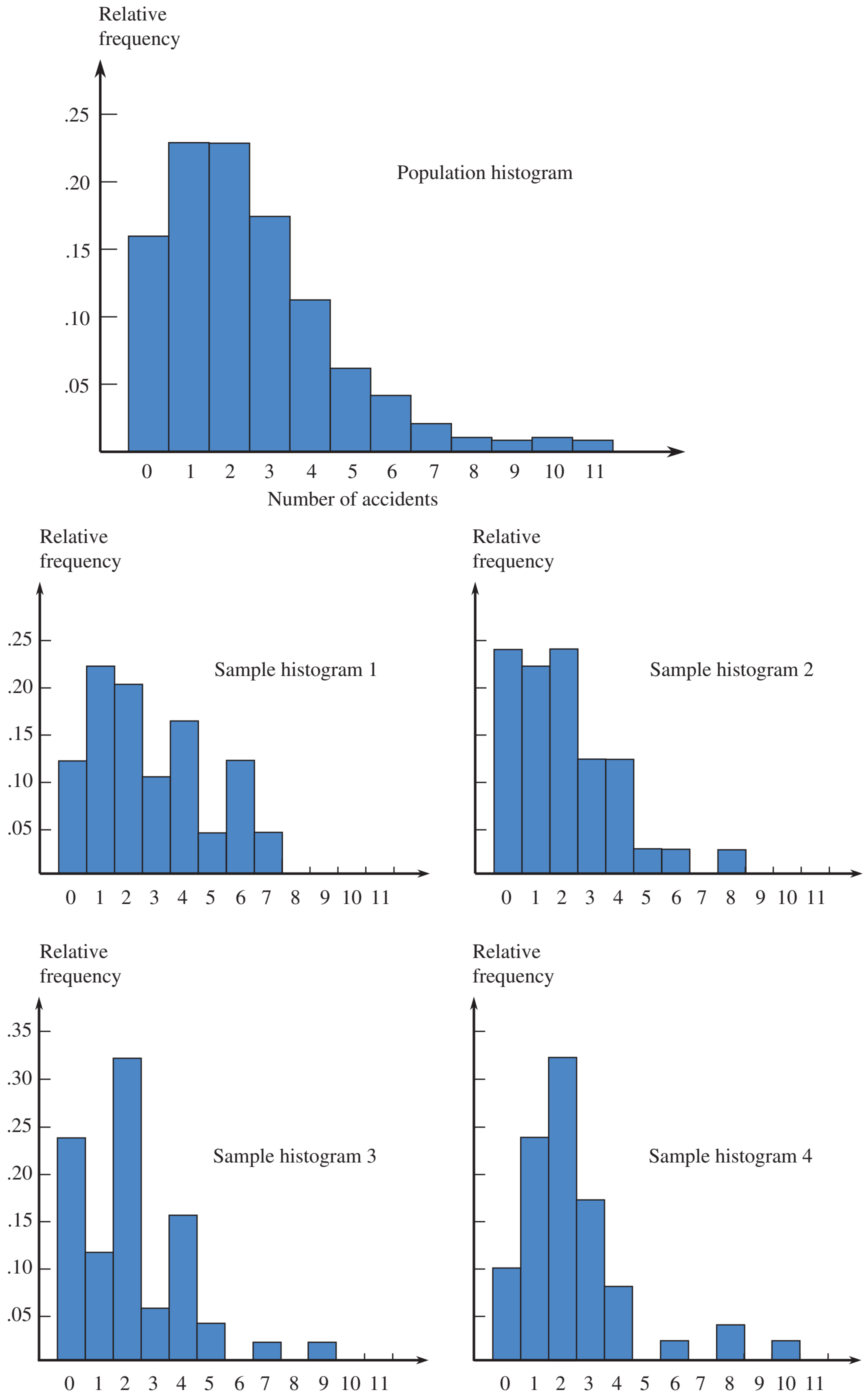


FIGURE 3.29
Comparison of population and sample histograms for number of accidents.

certainly resemble one another in a general way, but some dissimilarities are also obvious. The population histogram rises to a peak and then declines smoothly, whereas the sample histograms tend to have more peaks, valleys, and gaps. Although the population data set contained an observation of 11, none of the four samples did. In fact, in the first two samples, the largest observations were 7 and 8, respectively. In Chapters 8–15 we will see how sampling variability can be described and taken into account when we use sample data to draw conclusions about a population.

Cumulative Relative Frequencies and Cumulative Relative Frequency Plots

Rather than wanting to know what proportion of the data fall in a particular class, we often wish to determine the proportion falling below a specified value. This is easily done when the value is a class boundary.

Consider the following intervals and relative frequencies for carry-on luggage weight for passengers on flights between Phoenix and New York City during October 2009:

Class	0 to 5	5 to <10	10 to <15	15 to <20	. . .
Relative frequency	.05	.10	.18	.25	. . .

Then

$$\begin{aligned}&\text{proportion of passengers with carry-on luggage weight less than}\\&\text{15 lbs.} = \text{proportion in one of the first three classes}\\&\qquad = .05 + .10 + .18\\&\qquad = .33\end{aligned}$$

Similarly,

$$\begin{aligned}&\text{proportion of passengers with carry-on luggage weight less than}\\&\text{20 lbs.} = .05 + .10 + .18 + .25 = .33 + .25 = .58\end{aligned}$$

Each such sum of relative frequencies is called a **cumulative relative frequency**. Notice that the cumulative relative frequency .58 is the sum of the previous cumulative relative frequency .33 and the "current" relative frequency .25. The use of cumulative relative frequencies is illustrated in Example 3.19.

EXAMPLE 3.19 Albuquerque Rainfall

The National Climatic Data Center has been collecting weather data for many years. Annual rainfall totals for Albuquerque, New Mexico, from 1950 to 2008 (www.ncdc.noaa.gov/oa/climate/research/cag3/city.html) were used to construct the relative frequency distribution shown in Table 3.8. The table also contains a column of cumulative relative frequencies.

The proportion of years with annual rainfall less than 10 inches is .585, the cumulative relative frequency for the 9 to <10 interval. What about the proportion of years with annual rainfall less than 8.5 inches? Because 8.5 is not the endpoint of one of the intervals in the frequency distribution, we can only estimate this from the information given. The value 8.5 is halfway between the endpoints of the 8 to 9 inter-

TABLE 3.8 Relative Frequency distribution for Albuquerque Rainfall Data with Cumulative Relative Frequencies

Annual Rainfall (inches)	Frequency	Relative Frequency	Cumulative Relative Frequency
4 to <5	3	0.052	0.052
5 to <6	6	0.103	0.155 = .052 + .103
6 to <7	5	0.086	0.241 = .052 + .103 + .086 or .155 + .086
7 to <8	6	0.103	0.344
8 to <9	10	0.172	0.516
9 to <10	4	0.069	0.585
10 to <11	12	0.207	0.792
11 to <12	6	0.103	0.895
12 to <13	3	0.052	0.947
13 to <14	3	0.052	0.999

val, so it is reasonable to estimate that half of the relative frequency of .172 for this interval belongs in the 8 to 8.5 range. Then

$$\begin{pmatrix}\text{estimate of proportion of}\\ \text{years with rainfall less}\\ \text{than 8.5 inches}\end{pmatrix} = .052 + .103 + .086 + .103 + \frac{1}{2}(.172) = .430$$

This proportion could also have been computed using the cumulative relative frequencies as

$$\begin{pmatrix}\text{estimate of proportion of}\\ \text{years with rainfall less}\\ \text{than 8.5 inches}\end{pmatrix} = .344 + \frac{1}{2}(.172) = .430$$

Similarly, since 11.25 is one-fourth of the way between 11 and 12,

$$\begin{pmatrix}\text{estimate of proportion of}\\ \text{years with rainfall less}\\ \text{than 11.25 inches}\end{pmatrix} = .792 + \frac{1}{4}(.103) = .818$$

A **cumulative relative frequency** plot is just a graph of the cumulative relative frequencies against the upper endpoint of the corresponding interval. The pairs

(upper endpoint of interval, cumulative relative frequency)

are plotted as points on a rectangular coordinate system, and successive points in the plot are connected by a line segment. For the rainfall data of Example 3.19, the plotted points would be

(5, .052)	(6, .155)	(7, .241)	(8, .344)	(9, .516)
(10, .585)	(11, .792)	(12, .895)	(13, .947)	(14, .999)

One additional point, the pair (lower endpoint of first interval, 0), is also included in the plot (for the rainfall data, this would be the point (4 0)), and then points are connected by line segments. Figure 3.30 shows the cumulative relative frequency plot for the rainfall data. The cumulative relative frequency plot can be used to obtain approximate answers to questions such as

> What proportion of the observations is smaller than a particular value?

and

> What value separates the smallest *p* percent from the larger values?

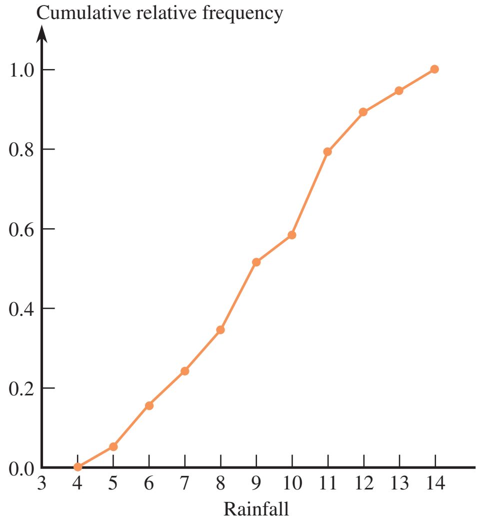


FIGURE 3.30
Cumulative relative frequency plot for the rainfall data of Example 3.19.

For example, to determine the approximate proportion of years with annual rainfall less than 9.5 inches, we would follow a vertical line up from 9.5 on the *x*-axis and then read across to the *y*-axis to obtain the corresponding relative frequency, as illustrated in Figure 3.31(a). Approximately .55, or 55%, of the years had annual rainfall less than 9.5 inches.

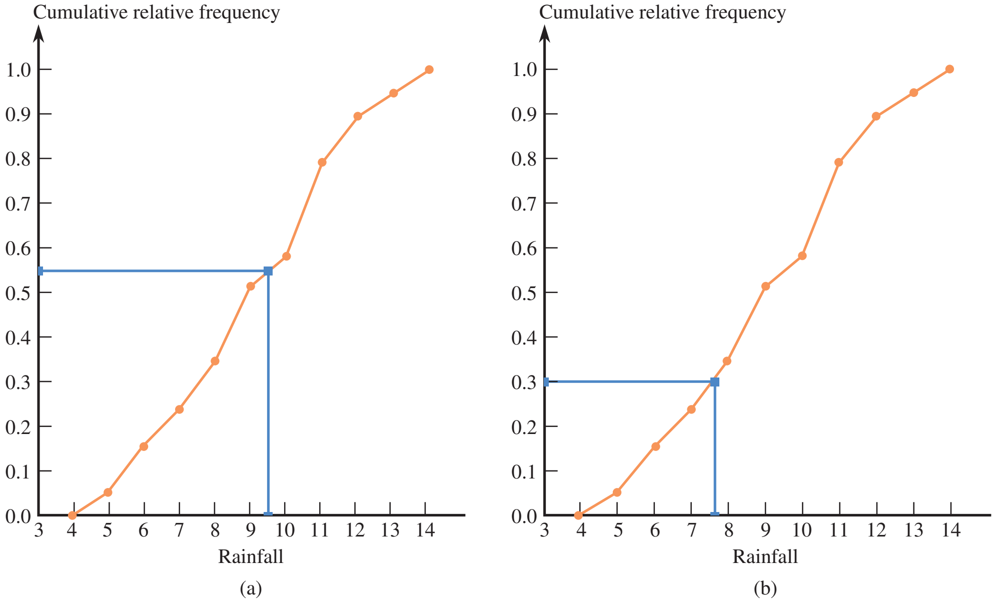


FIGURE 3.31
Using the cumulative relative frequency plot.
(a) Determining the approximate proportion of years with annual rainfall less than 9.5 inches.
(b) Finding the amount of rainfall that separates the 30% of years with the lowest rainfall from the 70% with higher rainfall.

Similarly, to find the amount of rainfall that separates the 30% of years with the smallest annual rainfall from years with higher rainfall, start at .30 on the cumulative relative frequency axis and move across and then down to find the corresponding rainfall amount, as shown in Figure 3.31(b). Approximately 30% of the years had annual rainfall of less than 7.6 inches.

EXERCISES 3.22 - 3.37

3.22 ● The article "Americans on the Move" (*USA Today*, November 30, 2007) included the data in the accompanying table. Entries in the table are the percentage of state residents who had moved during 2006.

State	Percentage of Residents Who Moved During 2006
Alabama	16.1
Alaska	21.2
Arizona	20.2
Arkansas	18.9
California	15.9
Colorado	19.6
Connecticut	13.1
Delaware	14.0
District of Columbia	18.8
Florida	17.4
Georgia	18.8
Hawaii	14.5
Idaho	21.0
Illinois	15.0
Indiana	16.8
Iowa	17.0
Kansas	18.7
Kentucky	16.8
Louisiana	18.9
Maine	14.4
Maryland	14.5
Massachusetts	13.6
Michigan	14.2
Minnesota	14.2
Mississippi	17.2
Missouri	17.5
Montana	17.5
Nebraska	18.0
Nevada	22.0
New Hampshire	13.7
New Jersey	11.1
New Mexico	16.8
New York	11.5
North Carolina	17.5
North Dakota	17.2
Ohio	15.7
Oklahoma	19.2
Oregon	20.2
Pennsylvania	12.7
Rhode Island	13.4
South Carolina	16.6
South Dakota	16.7
Tennessee	16.6
Texas	19.1
Utah	20.7
Vermont	14.5
Virginia	16.3
Washington	19.5
West Virginia	12.7
Wisconsin	15.3
Wyoming	18.8

(continued)

Construct a histogram of these data using class intervals of 10 to <12, 12 to <14, 14 to <16, and so on. Write a few sentences to describe the shape, center, and spread of the distribution.

3.23 ● The accompanying data on annual maximum wind speed (in meters per second) in Hong Kong for each year in a 45-year period were given in an article that appeared in the journal *Renewable Energy* (March, 2007). Use the annual maximum wind speed data to construct a histogram. Is the histogram approximately symmetric, positively skewed, or negatively skewed? Would you describe the histogram as unimodal, bimodal, or multimodal?

30.3	39.0	33.9	38.6	44.6	31.4	26.7	51.9	31.9
27.2	52.9	45.8	63.3	36.0	64.0	31.4	42.2	41.1
37.0	34.4	35.5	62.2	30.3	40.0	36.0	39.4	34.4
28.3	39.1	55.0	35.0	28.8	25.7	62.7	32.4	31.9
37.5	31.5	32.0	35.5	37.5	41.0	37.5	48.6	28.1

3.24 ● The accompanying relative frequency table is based on data from the 2007 College Bound Seniors Report for California (College Board, 2008).

Score on SAT Reasoning Exam	Relative Frequency for Males	Relative Frequency for Females
200 to <250	.0404	.0183
250 to < 300	.0546	.0299
300 to <350	.1076	.0700
350 to < 400	.1213	.0896
400 to < 450	.1465	.1286
450 to < 500	.1556	.1540
500 to < 550	.1400	.1667
550 to <600	.1126	.1550
600 to < 650	.0689	.1050
650 to < 700	.0331	.0529
700 to < 750	.0122	.0194
750 to <800	.0072	.0105

a. Construct a relative frequency histogram for SAT reasoning score for males.

b. Construct a relative frequency histogram for SAT reasoning score for females.

c. Based on the histograms from Parts (a) and (b), write a few sentences commenting on the similarities and differences in the distribution of SAT reasoning scores for males and females.

3.25 ● The data in the accompanying table represents the percentage of workers who are members of a union for each U.S. state and the District of Columbia (*AARP Bulletin*, September 2009).

State	% of Workers who Belong to a Union
Alabama	9.8
Alaska	23.5
Arizona	8.8
Arkansas	5.9
California	18.4
Colorado	8.0
Connecticut	16.9
Delaware	12.2
District of Columbia	13.4
Florida	6.4
Georgia	3.7
Hawaii	24.3
Idaho	7.1
Illinois	16.6
Indiana	12.4
Iowa	10.6
Kansas	7.0
Kentucky	8.6
Louisiana	4.6
Maine	12.3
Maryland	15.7
Massachusetts	12.6
Michigan	18.8
Minnesota	16.1
Mississippi	5.3
Missouri	11.2
Montana	12.2
Nebraska	8.3
Nevada	16.7
New Hampshire	3.5
New Jersey	6.1
New Mexico	10.6
New York	18.3
North Carolina	7.2
North Dakota	24.9
Ohio	14.2
Oklahoma	6.6
Oregon	16.6
Pennsylvania	15.4
Rhode Island	16.5
South Carolina	3.9
South Dakota	5.0
Tennessee	5.5
Texas	4.5
Utah	5.8
Vermont	4.1
Virginia	10.4
Washington	19.8
West Virginia	13.8
Wisconsin	15.0
Wyoming	7.7

(continued)

a. Construct a histogram of these data using class intervals of 0 to <5, 5 to <10, 10 to <15, 15 to <20, and 20 to <25.

b. Construct a dotplot of these data. Comment on the interesting features of the plot.

c. For this data set, which is a more informative graphical display—the dotplot from Part (b) or the histogram constructed in Part (a)? Explain.

d. Construct a histogram using about twice as many class intervals as the histogram in Part (a). Use 2.5 to <5 as the first class interval. Write a few sentences that explain why this histogram does a better job of displaying this data set than does the histogram in Part (a).

3.26 ● Medicare's new medical plans offer a wide range of variations and choices for seniors when picking a drug plan (*San Luis Obispo Tribune*, November 25, 2005). The monthly cost for a stand-alone drug plan varies from plan to plan and from state to state. The accompanying table gives the premium for the plan with the lowest cost for each state.

State	Cost per Month (dollars)
Alabama	14.08
Alaska	20.05
Arizona	6.14
Arkansas	10.31
California	5.41
Colorado	8.62
Connecticut	7.32
Delaware	6.44
District of Columbia	6.44
Florida	10.35
Georgia	17.91
Hawaii	17.18
Idaho	6.33
Illinois	13.32
Indiana	12.30
Iowa	1.87
Kansas	9.48
Kentucky	12.30
Louisiana	17.06
Maine	19.60
Maryland	6.44
Massachusetts	7.32
Michigan	13.75
Minnesota	1.87
Mississippi	11.60
Missouri	10.29
Montana	1.87
Nebraska	1.87
Nevada	6.42
New Hampshire	19.60
New Jersey	4.43
New Mexico	10.65
New York	4.10
North Carolina	13.27
North Dakota	1.87
Ohio	14.43
Oklahoma	10.07
Oregon	6.93
Pennsylvania	10.14
Rhode Island	7.32
South Carolina	16.57
South Dakota	1.87
Tennessee	14.08
Texas	10.31
Utah	6.33
Vermont	7.32
Virginia	8.81
Washington	6.93
West Virginia	10.14
Wisconsin	11.42
Wyoming	1.87

(continued)

a. Use class intervals of \$0 to <\$3, \$3 to <\$6, \$6 to <\$9, etc., to create a relative frequency distribution for these data.
b. Construct a histogram and comment on its shape.
c. Using the relative frequency distribution or the histogram, estimate the proportion of the states that have a minimum monthly plan of less than \$13.00 a month.

3.27 ● The following two relative frequency distributions were constructed using data that appeared in the report "Undergraduate Students and Credit Cards in 2004" (Nellie Mae, May 2005). One relative frequency distribution is based on credit bureau data for a random sample of 1413 college students, while the other is based on the result of a survey completed by 132 of the 1260 college students who received the survey.

Credit Card Balance (dollars)—Credit Bureau Data	Relative Frequency
0 to <100	.18
100 to <500	.19
500 to <1000	.14
1000 to <2000	.16
2000 to <3000	.10
3000 to <7000	.16
7000 or more	.07

Credit Card Balance (dollars)—Survey Data	Relative Frequency
0 to <100	.18
100 to <500	.22
500 to <1000	.17
1000 to <2000	.22
2000 to <3000	.07
3000 to <7000	.14
7000 or more	.00

a. Construct a histogram for the credit bureau data. For purposes of constructing the histogram, assume that none of the students in the sample had a balance higher than \$15,000 and that the last interval can be regarded as 7000 to <15,000. Be sure to use the density scale when constructing the histogram.

b. Construct a histogram for the survey data. Use the same scale that you used for the histogram in Part (a) so that it will be easy to compare the two histograms.

c. Comment on the similarities and differences in the histograms from Parts (a) and (b).

d. Do you think the high nonresponse rate for the survey may have contributed to the observed differences in the two histograms? Explain.

3.28 ● U.S. Census data for San Luis Obispo County, California, were used to construct the following frequency distribution for commute time (in minutes) of working adults (the given frequencies were read from a graph that appeared in the *San Luis Obispo Tribune* [September 1, 2002] and so are only approximate):

Commute Time	Frequency
0 to <5	5,200
5 to <10	18,200
10 to <15	19,600
15 to <20	15,400
20 to <25	13,800
25 to <30	5,700
30 to <35	10,200
35 to <40	2,000
40 to <45	2,000
45 to <60	4,000
60 to <90	2,100
90 to <120	2,200

a. Notice that not all intervals in the frequency distribution are equal in width. Why do you think that unequal width intervals were used?

b. Construct a table that adds a relative frequency and a density column to the given frequency distribution (see Example 3.17).

c. Use the densities computed in Part (b) to construct a histogram for this data set. (Note: The newspaper displayed an incorrectly drawn histogram based on frequencies rather than densities!) Write a few sentences commenting on the important features of the histogram.

d. Compute the cumulative relative frequencies, and construct a cumulative relative frequency plot.

e. Use the cumulative relative frequency plot constructed in Part (d) to answer the following questions.

   i. Approximately what proportion of commute times were less than 50 minutes?

   ii. Approximately what proportion of commute times were greater than 22 minutes?

   iii. What is the approximate commute time value that separates the shortest 50% of commute times from the longest 50%?

3.29 Student loans can add up, especially for those attending professional schools to study in such areas as medicine, law, or dentistry. Researchers at the University of Washington studied medical students and gave the following information on the educational debt of medical students on completion of their residencies (*Annals of Internal Medicine* [March 2002]: 384–398):

Educational Debt (dollars)	Relative Frequency
0 to <5000	.427
5000 to <20,000	.046
20,000 to <50,000	.109
50,000 to <100,000	.232
100,000 or more	.186

a. What are two reasons that you could not use the given information to construct a histogram with the educational debt intervals on the horizontal axis and relative frequency on the y-axis?

b. Suppose that no student had an educational debt of \$150,000 or more upon completion of his or her residency, so that the last class in the relative frequency distribution would be 100,000 to <150,000. Summarize this distribution graphically by constructing a histogram of the educational debt data. (Don't forget to use the density scale for the heights of the bars in the histogram, because the interval widths aren't all the same.)

c. Based on the histogram of Part (b), write a few sentences describing the educational debt of medical students completing their residencies.

3.30 An exam is given to students in an introductory statistics course. What is likely to be true of the shape of the histogram of scores if:

a. the exam is quite easy?

b. the exam is quite difficult?

c. half the students in the class have had calculus, the other half have had no prior college math courses, and the exam emphasizes mathematical manipulation?

Explain your reasoning in each case.

3.31 The accompanying frequency distribution summarizes data on the number of times smokers who had successfully quit smoking attempted to quit before their final successful attempt ("Demographic Variables, Smoking Variables, and Outcome Across Five Studies," *Health Psychology* [2007]: 278–287).

Number of Attempts	Frequency
0	778
1	306
2	274
3–4	221
5 or more	238

Assume that no one had made more than 10 unsuccessful attempts, so that the last entry in the frequency distribution can be regarded as 5–10 attempts. Summarize this data set using a histogram. Be careful—the class intervals are not all the same width, so you will need to use a density scale for the histogram. Also remember that for a discrete variable, the bar for 1 will extend from 0.5 to 1.5. Think about what this will mean for the bars for the 3–4 group and the 5–10 group.

3.32 ● Example 3.19 used annual rainfall data for Albuquerque, New Mexico, to construct a relative frequency distribution and cumulative relative frequency plot. The National Climate Data Center also gave the accompanying annual rainfall (in inches) for Medford, Oregon, from 1950 to 2008.

28.84	20.15	18.88	25.72	16.42	20.18	28.96	20.72	23.58	10.62
20.85	19.86	23.34	19.08	29.23	18.32	21.27	18.93	15.47	20.68
23.43	19.55	20.82	19.04	18.77	19.63	12.39	22.39	15.95	20.46
16.05	22.08	19.44	30.38	18.79	10.89	17.25	14.95	13.86	15.30
13.71	14.68	15.16	16.77	12.33	21.93	31.57	18.13	28.87	16.69
18.81	15.15	18.16	19.99	19.00	23.97	21.99	17.25	14.07	

a. Construct a relative frequency distribution for the Medford rainfall data.

b. Use the relative frequency distribution of Part (a) to construct a histogram. Describe the shape of the histogram.

c. Construct a cumulative relative frequency plot for the Medford rainfall data.

d. Use the cumulative relative frequency plot of Part (c) to answer the following questions:

   i. Approximately what proportion of years had annual rainfall less than 15.5 inches?

   ii. Approximately what proportion of years had annual rainfall less than 25 inches?

   iii. Approximately what proportion of years had annual rainfall between 17.5 and 25 inches?

3.33 The National Climate Data Center referenced in the previous exercise and Example 3.19 also gives rainfall data for a number of other U.S. cities. Go to the web site www.ncdc.noaa.gov/oa/climate/research/cag3/city.html and select one of the other cities. Use the data from 1950 to the most recent year for which data is available for the city you have selected to construct a relative frequency distribution and histogram. Write a few sentences comparing the distribution of annual rainfall values for the city you selected to the rainfall distribution for Medford, Oregon. (Use the histogram for Medford constructed in Exercise 3.32.)

3.34 The authors of the paper "Myeloma in Patients Younger than Age 50 Years Presents with More Favorable Features and Shows Better Survival" (*Blood* [2008]: 4039–4047) studied patients who had been diagnosed with stage 2 multiple myeloma prior to the age of 50. For each patient who received high dose chemotherapy, the number of years that the patient lived after the therapy (survival time) was recorded. The cumulative relative frequencies in the accompanying table were approximated from survival graphs that appeared in the paper.

Years Survived	Cumulative Relative Frequency
0 to <2	.10
2 to <4	.52
4 to <6	.54
6 to <8	.64
8 to <10	.68
10 to <12	.70
12 to <14	.72
14 to <16	1.00

a. Use the given information to construct a cumulative relative frequency plot.

b. Use the cumulative relative frequency plot from Part (a) to answer the following questions:

   i. What is the approximate proportion of patients who lived fewer than 5 years after treatment?

   ii. What is the approximate proportion of patients who lived fewer than 7.5 years after treatment?

   iii. What is the approximate proportion of patients who lived more than 10 years after treatment?

3.35

a. Use the cumulative relative frequencies given in the previous exercise to compute the relative frequencies for each class interval and construct a relative frequency distribution.

b. Summarize the survival time data with a histogram.

c. Based on the histogram, write a few sentences describing survival time of the stage 2 myeloma patients in this study.

d. What additional information would you need in order to decide if it is reasonable to generalize conclusions about survival time from the group of patients in the study to all patients younger than 50 years old who are diagnosed with multiple myeloma and who receive high dose chemotherapy?

3.36 Construct a histogram corresponding to each of the five frequency distributions, I–V, given in the following table, and state whether each histogram is symmetric, bimodal, positively skewed, or negatively skewed:

	Frequency				
Class Interval	I	II	III	IV	V
0 to <10	5	40	30	15	6
10 to <20	10	25	10	25	5
20 to <30	20	10	8	8	6
30 to <40	30	8	7	7	9
40 to <50	20	7	7	20	9
50 to <60	10	5	8	25	23
60 to <70	5	5	30	10	42

3.37 Using the five class intervals 100 to 120, 120 to 140, . . . , 180 to 200, devise a frequency distribution based on 70 observations whose histogram could be described as follows:

a. symmetric

b. bimodal

c. positively skewed

d. negatively skewed

3.4 Displaying Bivariate Numerical Data

A bivariate data set consists of measurements or observations on two variables, *x* and *y*. For example, *x* might be the distance from a highway and *y* the lead content of soil at that distance. When both *x* and *y* are numerical variables, each observation consists of a pair of numbers, such as (14, 5.2) or (27.63, 18.9). The first number in a pair is the value of *x*, and the second number is the value of *y*.

An unorganized list of bivariate data provides little information about the distribution of either the *x* values or the *y* values separately and even less information about how the two variables are related to one another. Just as graphical displays can be used to summarize univariate data, they can also help with bivariate data. The most important graph based on bivariate numerical data is a **scatterplot**.

In a scatterplot each observation (pair of numbers) is represented by a point on a rectangular coordinate system, as shown in Figure 3.32(a). The horizontal axis is identified with values of *x* and is scaled so that any *x* value can be easily located. Similarly, the vertical or *y*-axis is marked for easy location of *y* values. The point corresponding to any particular (*x*, *y*) pair is placed where a vertical line from the value on

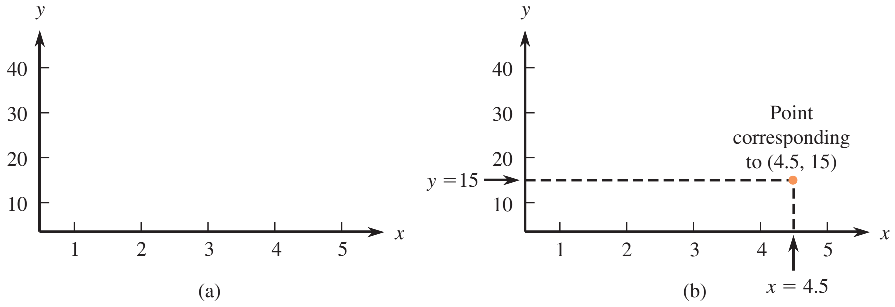


FIGURE 3.32
Constructing a scatterplot:
(a) rectangular coordinate system;
(b) point corresponding to (4.5, 15).

the *x*-axis meets a horizontal line from the value on the *y*-axis. Figure 3.32(b) shows the point representing the observation (4.5, 15); it is above 4.5 on the horizontal axis and to the right of 15 on the vertical axis.

EXAMPLE 3.20 Olympic Figure Skating

● Do tall skaters have an advantage when it comes to earning high artistic scores in figure skating competitions? Data on x = height (in cm) and y = artistic score in the free skate for both male and female singles skaters at the 2006 Winter Olympics are shown in the accompanying table. (Data set courtesy of John Walker.)

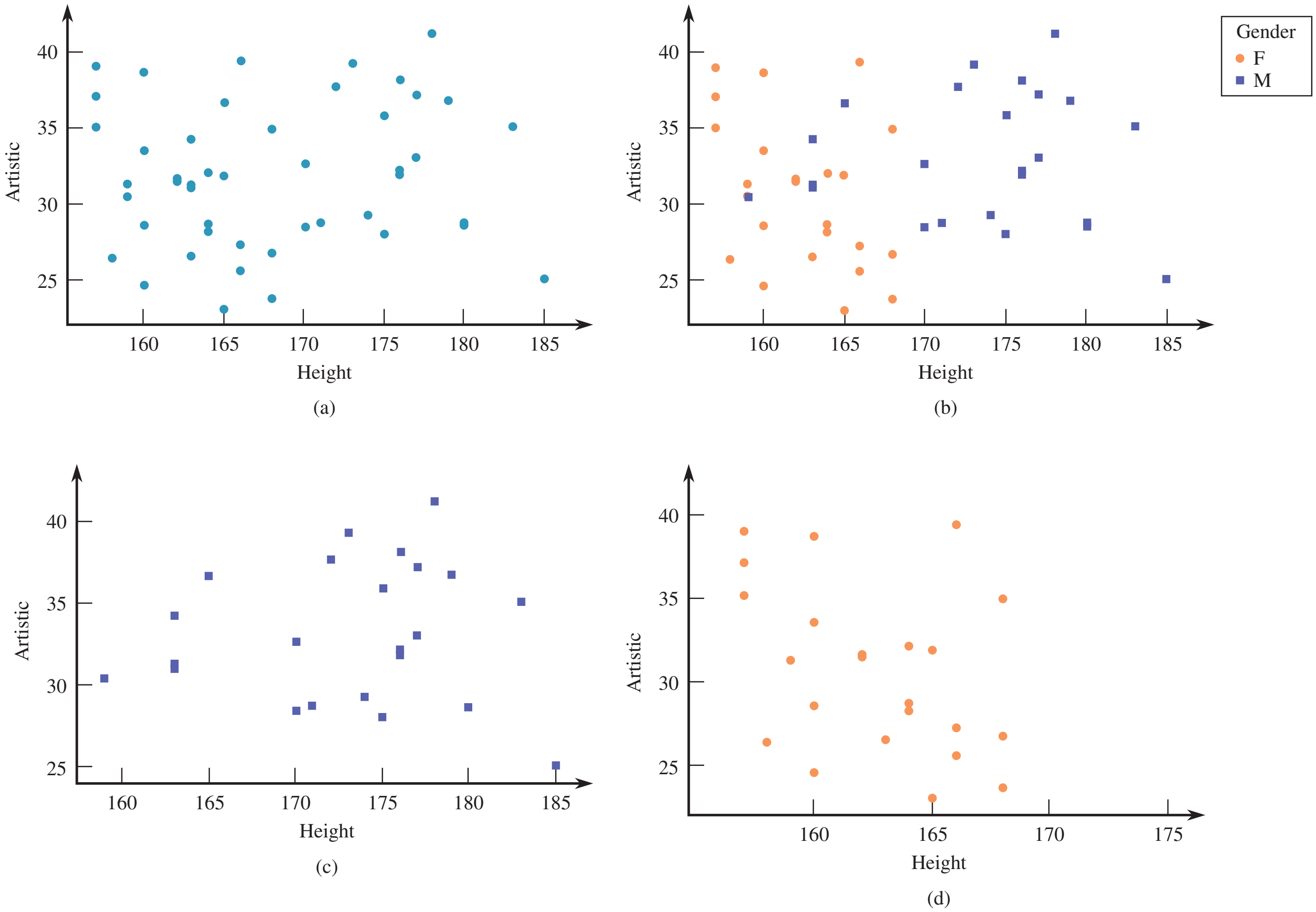
Name	Gender	Height	Artistic
PLUSHENKO Yevgeny	M	178	41.2100
BUTTLE Jeffrey	M	173	39.2500
LYSACEK Evan	M	177	37.1700
LAMBIEL Stephane	M	176	38.1400
SAVOIE Matt	M	175	35.8600
WEIR Johnny	M	172	37.6800
JOUBERT Brian	M	179	36.7900
VAN DER PERREN Kevin	M	177	33.0100
TAKAHASHI Daisuke	M	165	36.6500
KLIMKIN Ilia	M	170	32.6100
ZHANG Min	M	176	31.8600
SAWYER Shawn	M	163	34.2500
LI Chengjiang	M	170	28.4700
SANDHU Emanuel	M	183	35.1100
VERNER Tomas	M	180	28.6100
DAVYDOV Sergei	M	159	30.4700
CHIPER Gheorghe	M	176	32.1500
DINEV Ivan	M	174	29.2500
DAMBIER Frederic	M	163	31.2500
LINDEMANN Stefan	M	163	31.0000
KOVALEVSKI Anton	M	171	28.7500
BERNTSSON Kristoffer	M	175	28.0400
PFEIFER Viktor	M	180	28.7200
TOTH Zoltan	M	185	25.1000
ARAKAWA Shizuka	F	166	39.3750
COHEN Sasha	F	157	39.0063
SLUTSKAYA Irina	F	160	38.6688
SUGURI Fumie	F	157	37.0313
ROCHETTE Joannie	F	157	35.0813
MEISSNER Kimmie	F	160	33.4625
HUGHES Emily	F	165	31.8563
MEIER Sarah	F	164	32.0313
KOSTNER Carolina	F	168	34.9313
SOKOLOVA Yelena	F	162	31.4250
YAN Liu	F	164	28.1625
LEUNG Mira	F	168	26.7000
GEDEVANISHVILI Elene	F	159	31.2250
KORPI Kiira	F	166	27.2000
POYKIO Susanna	F	159	31.2125

● Data set available online

Name	Gender	Height	Artistic
ANDO Miki	F	162	31.5688
EFREMENKO Galina	F	163	26.5125
LIASHENKO Elena	F	160	28.5750
HEGEL Idora	F	166	25.5375
SEBESTYEN Julia	F	164	28.6375
KARADEMIR Tugba	F	165	23.0000
FONTANA Silvia	F	158	26.3938
PAVUK Viktoria	F	168	23.6688
MAXWELL Fleur	F	160	24.5438

Figure 3.33(a) gives a scatterplot of the data. Looking at the data and the scatterplot, we can see that

1. Several observations have identical *x* values but different *y* values (for example, $x = 176$ cm for both Stephane Lambiel and Min Zhang, but Lambiel's artistic score was 38.1400 and Zhang's artistic score was 31.8600). Thus, the value of *y* is *not* determined *solely* by the value of *x* but by various other factors as well.



(a)

(b)

(c)

(d)

FIGURE 3.33
Scatterplots for the data of Example 3.20: (a) scatterplot of data; (b) scatterplot of data with observations for males and females distinguished by color; (c) scatterplot for male skaters; (d) scatterplot for female skaters.

2. At any given height there is quite a bit of variability in artistic score. For example, for those skaters with height 160 cm, artistic scores ranged from a low of about 24.5 to a high of about 39.
3. There is no noticeable tendency for artistic score to increase as height increases. There does not appear to be a strong relationship between height and artistic score.

The data set used to construct the scatter plot included data for both male and female skaters. Figure 3.33(b) shows a scatterplot of the (height, artistic score) pairs with observations for male skaters shown in blue and observations for female skaters shown in orange. Not surprisingly, the female skaters tend to be shorter than the male skaters (the observations for females tend to be concentrated toward the left side of the scatterplot). Careful examination of this plot shows that while there was no apparent pattern in the combined (male and female) data set, there may be a relationship between height and artistic score for female skaters.

Figures 3.33(c) and 3.33(d) show separate scatterplots for the male and female skaters, respectively. It is interesting to note that it appears that for female skaters, higher artistic scores seem to be associated with smaller height values, but for men there does not appear to be a relationship between height and artistic score. The relationship between height and artistic score for women is not evident in the scatterplot of the combined data.

The horizontal and vertical axes in the scatterplots of Figure 3.33 do not intersect at the point (0, 0). In many data sets, the values of *x* or of *y* or of both variables differ considerably from 0 relative to the ranges of the values in the data set. For example, a study of how air conditioner efficiency is related to maximum daily outdoor temperature might involve observations at temperatures of 80°, 82°, . . . , 98°, 100°. In such cases, the plot will be more informative if the axes intersect at some point other than (0, 0) and are marked accordingly. This is illustrated in Example 3.21.

EXAMPLE 3.21 Taking Those "Hard" Classes Pays Off

● The report titled "2007 College Bound Seniors" (College Board, 2007) included the accompanying table showing the average score on the writing and math sections of the SAT for groups of high school seniors completing different numbers of years of study in six core academic subjects (arts and music, English, foreign languages, mathematics, natural sciences, and social sciences and history). Figure 3.34(a) and (b) show two scatterplots of x = total number of years of study and y = average writing SAT score. The scatterplots were produced by the statistical computer package Minitab. In Figure 3.34(a), we let Minitab select the scale for both axes. Figure 3.34(b) was obtained by specifying that the axes would intersect at the point (0, 0). The second plot does not make effective use of space. It is more crowded than the first plot, and such crowding can make it more difficult to see the general nature of any relationship. For example, it can be more difficult to spot curvature in a crowded plot.

Step-by-step technology instructions available online

● Data set available online

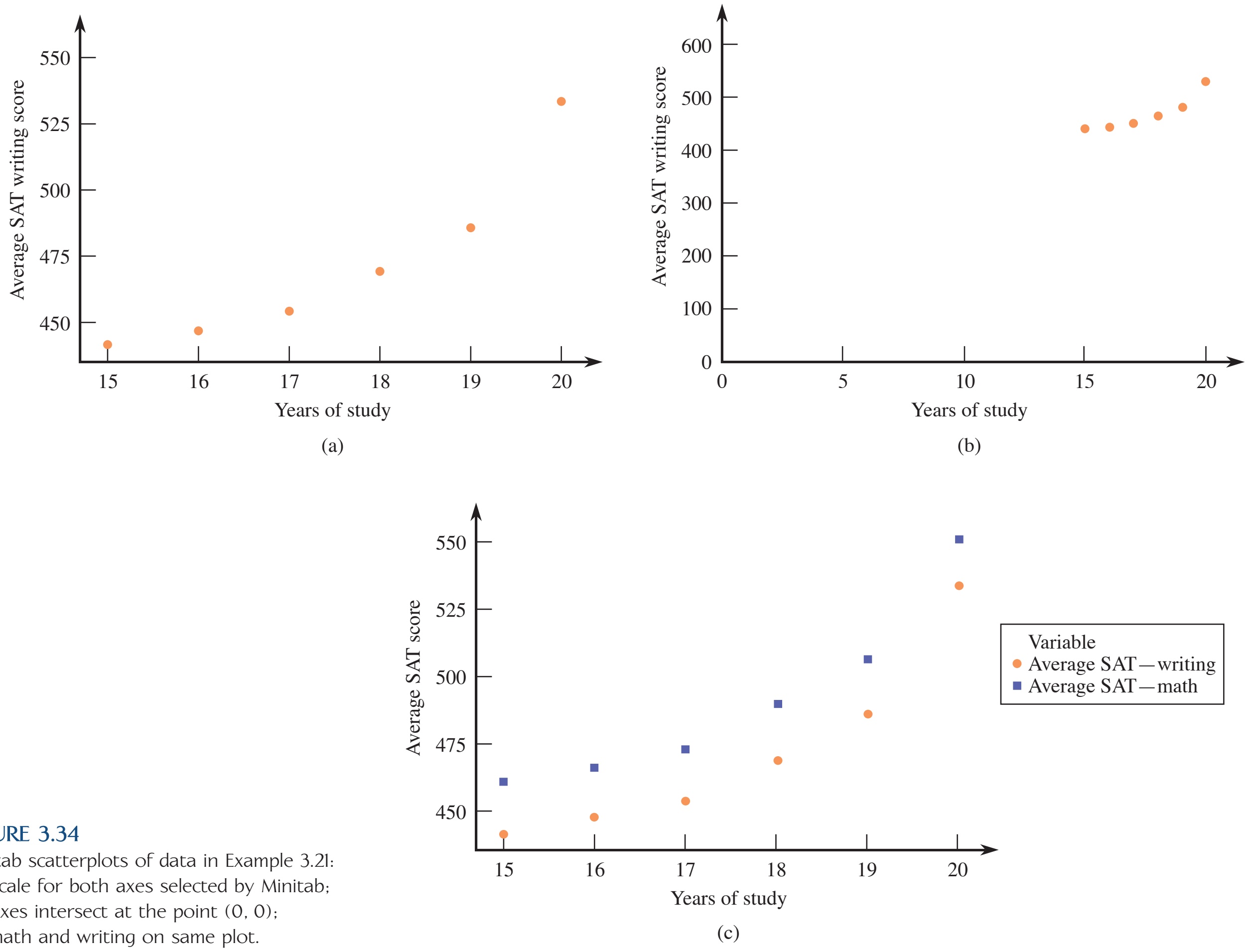


FIGURE 3.34
Minitab scatterplots of data in Example 3.21:
(a) scale for both axes selected by Minitab;
(b) axes intersect at the point (0, 0);
(c) math and writing on same plot.

Years of Study	Average Writing Score	Average Math Score
15	442	461
16	447	466
17	454	473
18	469	490
19	486	507
20	534	551

The scatterplot for average writing SAT score exhibits a fairly strong curved pattern, indicating that there is a strong relationship between average writing SAT score and the total number of years of study in the six core academic subjects. Although the pattern in the plot is curved rather than linear, it is still easy to see that the average writing SAT score increases as the number of years of study increases. Figure 3.34(c) shows a scatterplot with the average writing SAT scores represented by blue squares and the average math SAT scores represented by orange dots. From this plot we can see that while the average math SAT scores tend to be higher than the average writing scores at all of the values of total number of years of study, the general curved form of the relationship is similar.

In Chapter 5, methods for summarizing bivariate data when the scatterplot reveals a pattern are introduced. Linear patterns are relatively easy to work with. A curved pattern, such as the one in Example 3.21, is a bit more complicated to analyze, and methods for summarizing such nonlinear relationships are developed in Section 5.4.

Time Series Plots

Data sets often consist of measurements collected over time at regular intervals so that we can learn about change over time. For example, stock prices, sales figures, and other socio-economic indicators might be recorded on a weekly or monthly basis. A **time-series plot** (sometimes also called a time plot) is a simple graph of data collected over time that can be invaluable in identifying trends or patterns that might be of interest.

A time-series plot can be constructed by thinking of the data set as a bivariate data set, where y is the variable observed and x is the time at which the observation was made. These (x, y) pairs are plotted as in a scatterplot. Consecutive observations are then connected by a line segment; this aids in spotting trends over time.

EXAMPLE 3.22 The Cost of Christmas

The Christmas Price Index is computed each year by PNC Advisors, and it is a humorous look at the cost of the giving all of the gifts described in the popular Christmas song "The Twelve Days of Christmas." The year 2008 was the most costly year since the index began in 1984, with the "cost of Christmas" at $21,080. A plot of the Christmas Price Index over time appears on the PNC web site (www.pncchristmaspriceindex.com) and the data given there were used to construct the time-series plot of Figure 3.35. The plot shows an upward trend in the index from

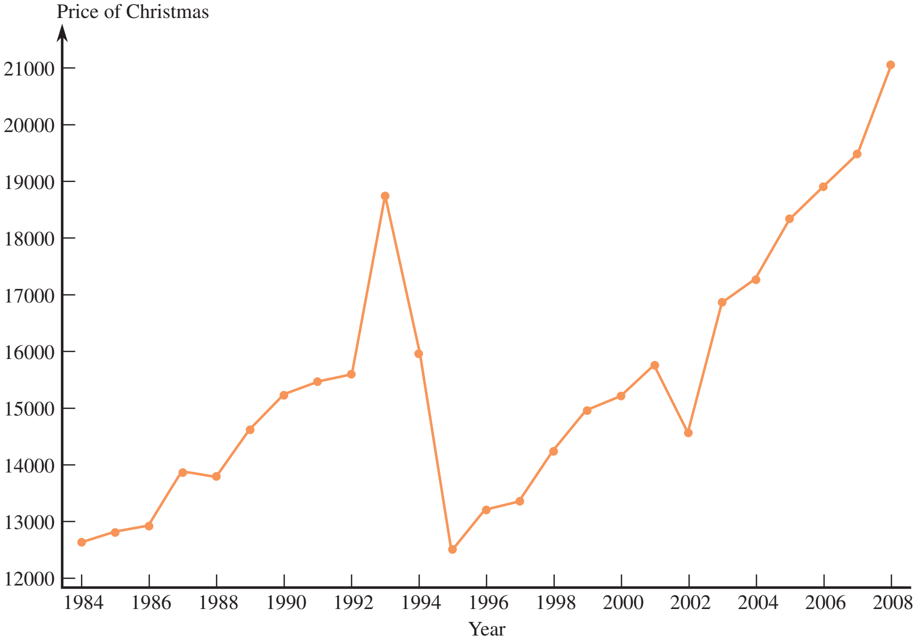


FIGURE 3.35
Time-series plot for the Christmas Price Index data of Example 3.22.

1984 until 1993. A dramatic drop in the cost occurred between 1993 and 1995, but there has been a clear upward trend in the index since then. You can visit the web site to see individual time-series plots for each of the twelve gifts that are used to determine the Christmas Price Index (a partridge in a pear tree, two turtle doves, etc.). See if you can figure out what caused the dramatic decline in 1995.

EXAMPLE 3.23 Education Level and Income—Stay in School!

The time-series plot shown in Figure 3.36 appears on the U.S. Census Bureau web site. It shows the average earnings of workers by educational level as a proportion of the average earnings of a high school graduate over time. For example, we can see from this plot that in 1993 the average earnings for people with bachelor's degrees was about 1.5 times the average for high school graduates. In that same year, the average earnings for those who were not high school graduates was only about 75% (a proportion of .75) of the average for high school graduates. The time-series plot also shows that the gap between the average earnings for high school graduates and those with a bachelor's degree or an advanced degree widened during the 1990s.

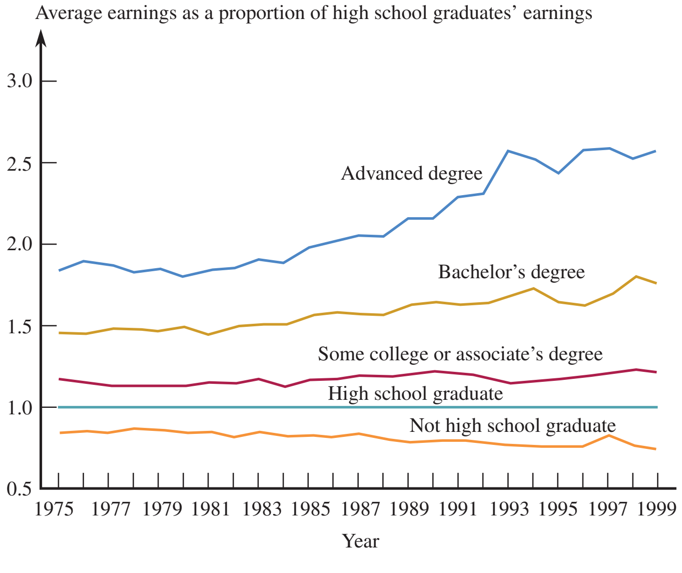


FIGURE 3.36 Time-series plot for average earnings as a proportion of the average earnings of high school graduates.

EXERCISES 3.38 - 3.45

3.38 ● *Consumer Reports Health* (www.consumerreports.org) gave the accompanying data on saturated fat (in grams), sodium (in mg), and calories for 36 fast-food items.

Fat	Sodium	Calories
2	1042	268
5	921	303
3	250	260
2	770	660
1	635	180
6	440	290
4.5	490	290
5	1160	360
3.5	970	300
1	1120	315
2	350	160
3	450	200
6	800	320
3	1190	420
2	1090	120
5	570	290
3.5	1215	285
2.5	1160	390
0	520	140
2.5	1120	330
1	240	120
3	650	180
1	1620	340
4	660	380
3	840	300
1.5	1050	490
3	1440	380
9	750	560
1	500	230
1.5	1200	370
2.5	1200	330
3	1250	330
0	1040	220
0	760	260
2.5	780	220
3	500	230

a. Construct a scatterplot using y = calories and x = fat. Does it look like there is a relationship between fat and calories? Is the relationship what you expected? Explain.

b. Construct a scatterplot using y = calories and x = sodium. Write a few sentences commenting on the difference between the relationship of calories to fat and calories to sodium.

c. Construct a scatterplot using y = sodium and x = fat. Does there appear to be a relationship between fat and sodium?

d. Add a vertical line at $x = 3$ and a horizontal line at $y = 900$ to the scatterplot in Part (c). This divides the scatterplot into four regions, with some of the points in the scatterplot falling into each of the four regions. Which of the four regions corresponds to healthier fast-food choices? Explain.

3.39 The report "Wireless Substitution: Early Release of Estimates from the National Health Interview Survey" (Center for Disease Control, 2009) gave the following estimates of the percentage of homes in the United States that had only wireless phone service at 6-month intervals from June 2005 to December 2008.

Date	Percent with Only Wireless Phone Service
June 2005	7.3
December 2005	8.4
June 2006	10.5
December 2006	12.8
June 2007	13.6
December 2007	15.8
June 2008	17.5
December 2008	20.2

Construct a time-series plot for these data and describe the trend in the percent of homes with only wireless phone service over time. Has the percent increased at a fairly steady rate?

3.40 ● The accompanying table gives the cost and an overall quality rating for 15 different brands of bike helmets (www.consumerreports.org).

Cost	Rating
35	65
20	61
30	60
40	55
50	54
23	47
30	47
18	43
40	42
28	41
20	40

(continued)

Bold exercises answered in back ● Data set available online ✦ Video Solution available

Cost	Rating
25	32
30	63
30	63
40	53

a. Construct a scatterplot using y = quality rating and x = cost.

b. Based on the scatterplot from Part (a), does there appear to be a relationship between cost and quality rating? Does the scatterplot support the statement that the more expensive bike helmets tended to receive higher quality ratings?

3.41 ● The accompanying table gives the cost and an overall quality rating for 10 different brands of men's athletic shoes and nine different brands of women's athletic shoes (www.consumerreports.org).

Cost	Rating	Type
65	71	Men's
45	70	Men's
45	62	Men's
80	59	Men's
110	58	Men's
110	57	Men's
30	56	Men's
80	52	Men's
110	51	Men's
70	51	Men's
65	71	Women's
70	70	Women's
85	66	Women's
80	66	Women's
45	65	Women's
70	62	Women's
55	61	Women's
110	60	Women's
70	59	Women's

a. Using the data for all 19 shoes, construct a scatterplot using y = quality rating and x = cost. Write a sentence describing the relationship between quality rating and cost.

b. Construct a scatterplot of the 19 data points that uses different colors or different symbols to distinguish the points that correspond to men's shoes from those that correspond to women's shoes. How do men's and women's athletic shoes differ with respect to cost and quality rating? Are the relationships between cost and quality rating the same for men and women? If not, how do the relationships differ?

3.42 ● The article "Medicine Cabinet is a Big Killer" (*The Salt Lake Tribune,* August 1, 2007) looked at the number of prescription-drug-overdose deaths in Utah over the period from 1991 to 2006. Construct a time-series plot for these data and describe the trend over time. Has the number of overdose deaths increased at a fairly steady rate?

Year	Number of Overdose Deaths
1991	32
1992	52
1993	73
1994	61
1995	68
1996	64
1997	85
1998	89
1999	88
2000	109
2001	153
2002	201
2003	237
2004	232
2005	308
2006	307

3.43 ● The article "Cities Trying to Rejuvenate Recycling Efforts" (*USA Today,* October 27, 2006) states that the amount of waste collected for recycling has grown slowly in recent years. This statement was supported by the data in the accompanying table. Use these data to construct a time-series plot. Explain how the plot is or is not consistent with the given statement.

Year	Recycled Waste (in millions of tons)
1990	29.7
1991	32.9
1992	36.0
1993	37.9
1994	43.5
1995	46.1
1996	46.4
1997	47.3
1998	48.0
1999	50.1
2000	52.7
2001	52.8
2002	53.7
2003	55.8
2004	57.2
2005	58.4

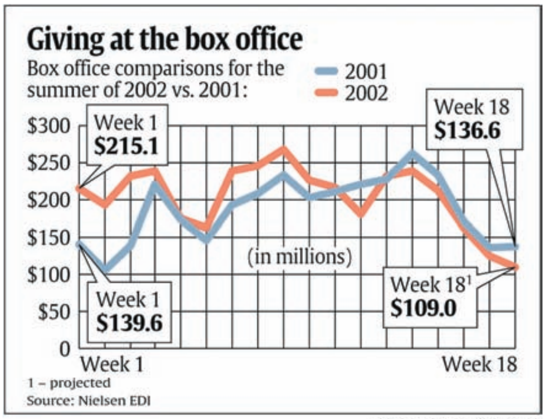
3.44 ● Some days of the week are more dangerous than others, according to Traffic Safety Facts produced by the National Highway Traffic Safety Administration. The average number of fatalities per day for each day of the week are shown in the accompanying table.

Average Fatalities per Day (day of the week)

	Mon	Tue	Wed	Thurs	Fri	Sat	Sun
1978–1982	103	101	107	116	156	201	159
1983–1987	98	96	99	108	140	174	140
1988–1992	97	94	97	106	139	168	135
1993–1997	97	93	96	102	129	148	127
1998–2002	99	96	98	104	129	149	130
Total	99	96	100	107	138	168	138

a. Using the midpoint of each year range (e.g., 1980 for the 1978–1982 range), construct a time-series plot that shows the average fatalities over time for each day of the week. Be sure to label each line clearly as to which day of the week it represents.
b. Write a sentence or two commenting on the difference in average number of fatalities for the days of the week. What is one possible reason for the differences?
c. Write a sentence or two commenting on the change in average number of fatalities over time. What is one possible reason for the change?

3.45 The accompanying time-series plot of movie box office totals (in millions of dollars) over 18 weeks of summer for both 2001 and 2002 appeared in *USA Today* (September 3, 2002):



By Karl Gelles, USA TODAY

USA TODAY. September 03, 2002. Reprinted with permission.

Patterns that tend to repeat on a regular basis over time are called seasonal patterns. Describe any seasonal patterns that you see in the summer box office data. Hint: Look for patterns that seem to be consistent from year to year.

Bold exercises answered in back ● Data set available online ✦ Video Solution available

3.5 Interpreting and Communicating the Results of Statistical Analyses

A graphical display, when used appropriately, can be a powerful tool for organizing and summarizing data. By sacrificing some of the detail of a complete listing of a data set, important features of the data distribution are more easily seen and more easily communicated to others.

Communicating the Results of Statistical Analyses

When reporting the results of a data analysis, a good place to start is with a graphical display of the data. A well-constructed graphical display is often the best way to highlight the essential characteristics of the data distribution, such as shape and spread for numerical data sets or the nature of the relationship between the two variables in a bivariate numerical data set.

For effective communication with graphical displays, some things to remember are

- Be sure to select a display that is appropriate for the given type of data.
- Be sure to include scales and labels on the axes of graphical displays.
- In comparative plots, be sure to include labels or a legend so that it is clear which parts of the display correspond to which samples or groups in the data set.

- Although it is sometimes a good idea to have axes that do not cross at (0, 0) in a scatterplot, the vertical axis in a bar chart or a histogram should always start at 0 (see the cautions and limitations later in this section for more about this).
- Keep your graphs simple. A simple graphical display is much more effective than one that has a lot of extra "junk." Most people will not spend a great deal of time studying a graphical display, so its message should be clear and straightforward.
- Keep your graphical displays honest. People tend to look quickly at graphical displays, so it is important that a graph's first impression is an accurate and honest portrayal of the data distribution. In addition to the graphical display itself, data analysis reports usually include a brief discussion of the features of the data distribution based on the graphical display.
- For categorical data, this discussion might be a few sentences on the relative proportion for each category, possibly pointing out categories that were either common or rare compared to other categories.
- For numerical data sets, the discussion of the graphical display usually summarizes the information that the display provides on three characteristics of the data distribution: center or location, spread, and shape.
- For bivariate numerical data, the discussion of the scatterplot would typically focus on the nature of the relationship between the two variables used to construct the plot.
- For data collected over time, any trends or patterns in the time-series plot would be described.

Interpreting the Results of Statistical Analyses

When someone uses a web search engine, do they rely on the ranking of the search results returned or do they first scan the results looking for the most relevant? The authors of the paper "Learning User Interaction Models for Predicting Web Search Result Preferences" (*Proceedings of the 29th Annual ACM Conference on Research and Development in Information Retrieval,* 2006) attempted to answer this question by observing user behavior when they varied the position of the most relevant result in the list of resources returned in response to a web search. They concluded that people clicked more often on results near the top of the list, even when they were not relevant. They supported this conclusion with the comparative bar graph in Figure 3.37.

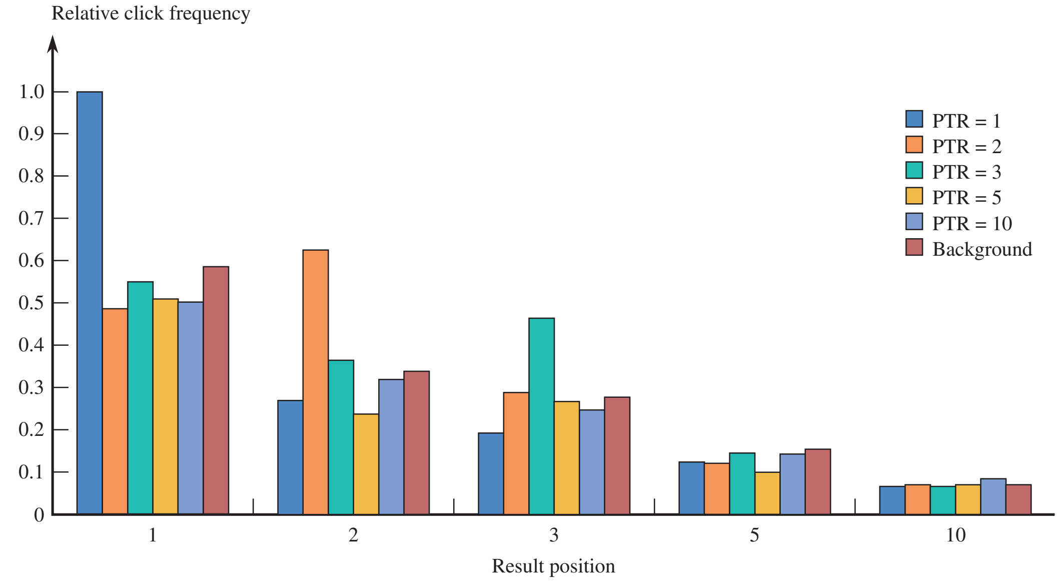


FIGURE 3.37 Comparative bar graph for click frequency data.

Although this comparative bar chart is a bit complicated, we can learn a great deal from this graphical display. Let's start by looking at the first group of bars. The different bars correspond to where in the list of search results the result that was considered to be most relevant was located. For example, in the legend PTR = 1 means that the most relevant result was in position 1 in the list returned. PTR = 2 means that the most relevant result was in the second position in the list returned, and so on. PTR = Background means that the most relevant result was not in the first 10 results returned. The first group of bars shows the proportion of times users clicked on the first result returned. Notice that all users clicked on the first result when it was the most relevant, but nearly half clicked on the first result when the most relevant result was in the second position and more than half clicked on the first result when the most relevant result was even farther down the list.

The second group of bars represents the proportion of users who clicked on the second result. Notice that the proportion who clicked on the second result was highest when the most relevant result was in that position. Stepping back to look at the entire graphical display, we see that users tended to click on the most relevant result if it was in one of the first three positions, but if it appeared after that, very few selected it. Also, if the most relevant result was in the third or a later position, users were more likely to click on the first result returned, and the likelihood of a click on the most relevant result decreased the farther down the list it appeared. To fully understand why the researchers' conclusions are justified, we need to be able to extract this kind of information from graphical displays.

The use of graphical data displays is quite common in newspapers, magazines, and journals, so it is important to be able to extract information from such displays. For example, data on test scores for a standardized math test given to eighth graders in 37 states, 2 territories (Guam and the Virgin Islands), and the District of Columbia were used to construct the stem-and-leaf display and histogram shown in Figure 3.38. Careful examination of these displays reveals the following:

1. Most of the participating states had average eighth-grade math scores between 240 and 280. We would describe the shape of this display as negatively skewed, because of the longer tail on the low end of the distribution.
2. Three of the average scores differed substantially from the others. These turn out to be 218 (Virgin Islands), 229 (District of Columbia), and 230 (Guam). These

21H	8
22L	
22H	9
23L	0
23H	
24L	
24H	79
25L	014
25H	6667779999
26L	0003344
26H	55778
27L	12233
27H	667
28L	01

Stem: Tens
Leaf: Ones

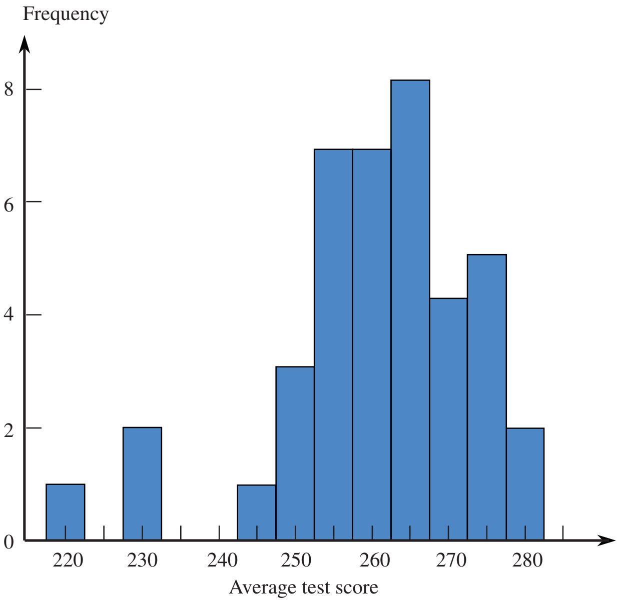


FIGURE 3.38 Stem-and-leaf display and histogram for math test scores.

three scores could be described as outliers. It is interesting to note that the three unusual values are from the areas that are not states.

3. There do not appear to be any outliers on the high side.
4. A "typical" average math score for the 37 states would be somewhere around 260.
5. There is quite a bit of variability in average score from state to state.

How would the displays have been different if the two territories and the District of Columbia had not participated in the testing? The resulting histogram is shown in Figure 3.39. Note that the display is now more symmetric, with no noticeable outliers. The display still reveals quite a bit of state-to-state variability in average score, and 260 still looks reasonable as a "typical" average score. Now suppose that the two highest values among the 37 states (Montana and North Dakota) had been even higher. The stem-and-leaf display might then look like the one given in Figure 3.40. In this stem-and-leaf display, two values stand out from the main part of the display. This would catch our attention and might cause us to look carefully at these two states to determine what factors may be related to high math scores.

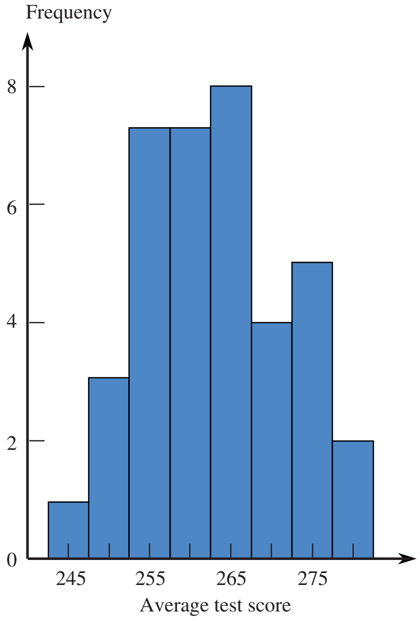


FIGURE 3.39
Histogram frequency for the modified math score data.

24H	79
25L	014
25H	6667779999
26L	0003344
26H	55778
27L	12233
27H	667
28L	
28H	
29L	
29H	68

Stem: Tens
Leaf: Ones

FIGURE 3.40
Stem-and-leaf display for modified math score data.

What to Look for in Published Data

Here are some questions you might ask yourself when attempting to extract information from a graphical data display:

- Is the chosen display appropriate for the type of data collected?
- For graphical displays of univariate numerical data, how would you describe the shape of the distribution, and what does this say about the variable being summarized?
- Are there any outliers (noticeably unusual values) in the data set? Is there any plausible explanation for why these values differ from the rest of the data? (The presence of outliers often leads to further avenues of investigation.)
- Where do most of the data values fall? What is a typical value for the data set? What does this say about the variable being summarized?
- Is there much variability in the data values? What does this say about the variable being summarized?

Of course, you should always think carefully about how the data were collected. If the data were not gathered in a reasonable manner (based on sound sampling methods or experimental design principles), you should be cautious in formulating any conclusions based on the data.

Consider the histogram in Figure 3.41, which is based on data published by the National Center for Health Statistics. The data set summarized by this histogram consisted of infant mortality rates (deaths per 1000 live births) for the 50 states in the United States. A histogram is an appropriate way of summarizing these data (although with only 50 observations, a stem-and-leaf display would also have been reasonable). The histogram itself is slightly positively skewed, with most mortality rates between 7.5 and 12. There is quite a bit of variability in infant mortality rate from state to state—perhaps more than we might have expected. This variability might be explained by differences in economic conditions or in access to health care. We may want to look further into these issues. Although there are no obvious outliers, the upper tail is a little longer than the lower tail. The three largest values in the data set are 12.1 (Alabama), 12.3 (Georgia), and 12.8 (South Carolina)—all Southern states. Again, this may suggest some interesting questions that deserve further investigation. A typical infant mortality rate would be about 9.5 deaths per 1000 live births. This represents an improvement, because researchers at the National Center for Health Statistics stated that the overall rate for 1988 was 10 deaths per 1000 live births. However, they also point out that the United States still ranked 22 out of 24 industrialized nations surveyed, with only New Zealand and Israel having higher infant mortality rates.

A Word to the Wise: Cautions and Limitations

When constructing and interpreting graphical displays, you need to keep in mind these things:

1. *Areas should be proportional to frequency, relative frequency, or magnitude of the number being represented.* The eye is naturally drawn to large areas in graphical displays, and it is natural for the observer to make informal comparisons based

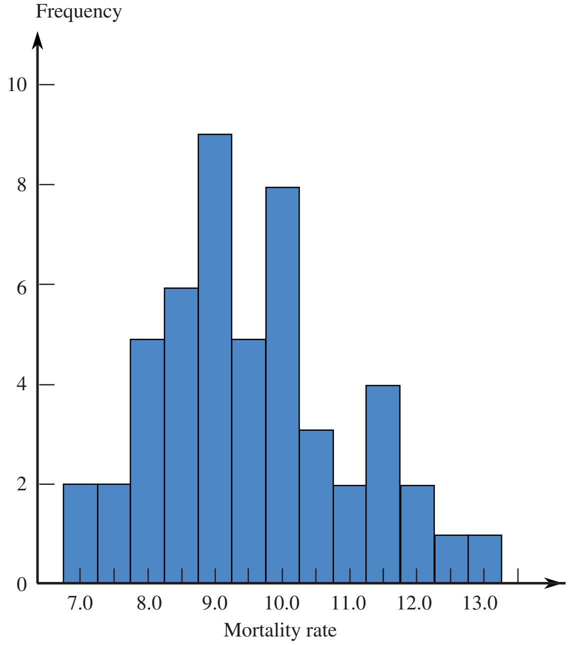
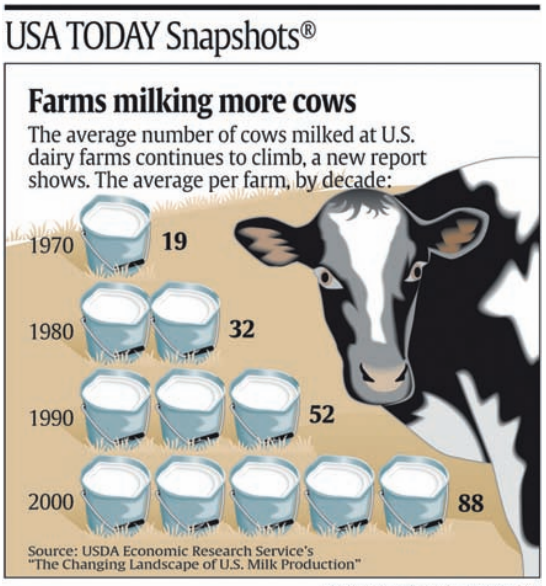


FIGURE 3.41
Histogram of infant mortality rates.

on area. Correctly constructed graphical displays, such as pie charts, bar charts, and histograms, are designed so that the areas of the pie slices or the bars are proportional to frequency or relative frequency. Sometimes, in an effort to make graphical displays more interesting, designers lose sight of this important principle, and the resulting graphs are misleading. For example, consider the following graph (*USA Today*, October 3, 2002):



By Marcy E. Mullins, USA TODAY

USA TODAY. October 03, 2002. Reprinted with permission.

In trying to make the graph more visually interesting by replacing the bars of a bar chart with milk buckets, areas are distorted. For example, the two buckets for 1980 represent 32 cows, whereas the one bucket for 1970 represents 19 cows. This is misleading because 32 is not twice as big as 19. Other areas are distorted as well.

Another common distortion occurs when a third dimension is added to bar charts or pie charts. For example, the pie chart at the bottom left of the page appeared in *USA Today* (September 17, 2009).

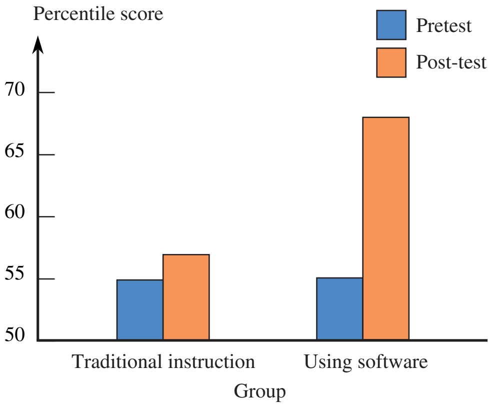
Adding the third dimension distorts the areas and makes it much more difficult to interpret correctly. A correctly drawn pie chart is shown below.

Image not available due to copyright restrictions

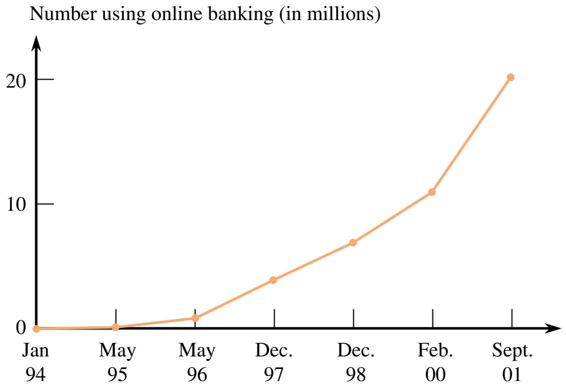
Image not available due to copyright restrictions

2. *Be cautious of graphs with broken axes.* Although it is common to see scatterplots with broken axes, be extremely cautious of time-series plots, bar charts, or histograms with broken axes. The use of broken axes in a scatterplot does not distort information about the nature of the relationship in the bivariate data set used to construct the display. On the other hand, in time-series plots, broken axes can sometimes exaggerate the magnitude of change over time. Although it is not always inadvisable to break the vertical axis in a time-series plot, it is something you should watch for, and if you see a time-series plot with a broken axis, as in the accompanying time-series plot of mortgage rates (*USA Today,* October 25, 2002), you should pay particular attention to the scale on the vertical axis and take extra care in interpreting the graph.

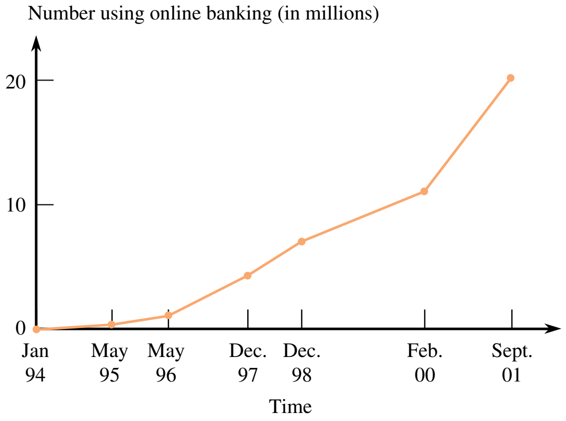
   In bar charts and histograms, the vertical axis (which represents frequency, relative frequency, or density) should *never* be broken. If the vertical axis is broken in this type of graph, the resulting display will violate the "proportional area" principle and the display will be misleading. For example, the accompanying bar chart is similar to one appearing in an advertisement for a software product designed to help teachers raise student test scores. By starting the vertical axis at 50, the gain for students using the software is exaggerated. Areas of the bars are not proportional to the magnitude of the numbers represented—the area for the rectangle representing 68 is more than three times the area of the rectangle representing 55!



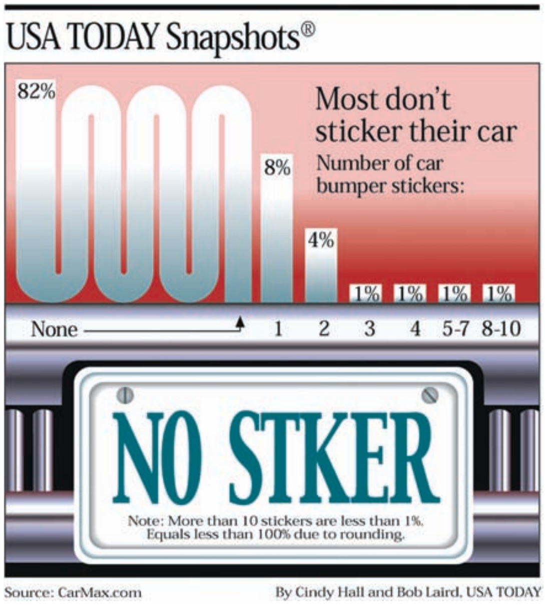
3. *Watch out for unequal time spacing in time-series plots.* If observations over time are not made at regular time intervals, special care must be taken in constructing the time-series plot. Consider the accompanying time-series plot, which is similar to one appearing in the *San Luis Obispo Tribune* (September 22, 2002) in an article on online banking:



Notice that the intervals between observations are irregular, yet the points in the plot are equally spaced along the time axis. This makes it difficult to make a coherent assessment of the rate of change over time. This could have been remedied by spacing the observations differently along the time axis, as shown in the following plot:



4. *Be careful how you interpret patterns in scatterplots.* A strong pattern in a scatterplot means that the two variables tend to vary together in a predictable way, but it does not mean that there is a cause-and-effect relationship between the two variables. We will consider this point further in Chapter 5, but in the meantime, when describing patterns in scatterplots, be careful not to use wording that implies that changes in one variable *cause* changes in the other.
5. *Make sure that a graphical display creates the right first impression.* For example, consider the graph below from *USA Today* (June 25, 2002). Although this graph does not violate the proportional area principle, the way the "bar" for the "none" category is displayed makes this graph difficult to read, and a quick glance at this graph would leave the reader with an incorrect impression.



USA TODAY. June 25, 2002. Used with permission.

EXERCISES 3.46 - 3.51

Image not available due to copyright restrictions

Suppose that you plan to include this graph in an article that you are writing for your school newspaper. Write a few paragraphs that could accompany the graph. Be sure to address what the graph reveals about how teen cell phone ownership is related to age and how it has changed over time.

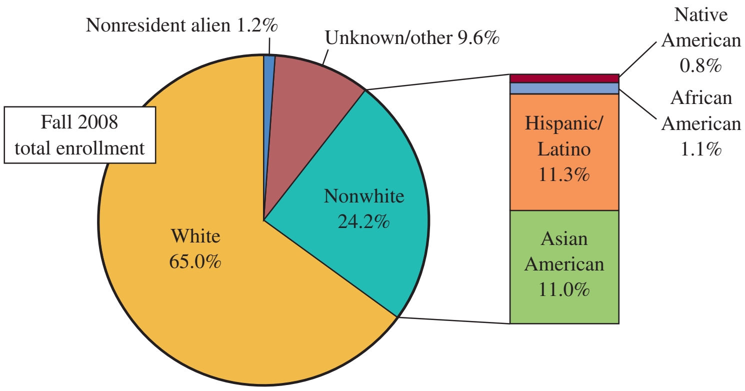


FIGURE EX-3.47

3.47 Figure EX-3.47 is from the Fall 2008 Census Enrollment Report at Cal Poly, San Luis Obispo. It uses both a pie chart and a segmented bar graph to summarize data on ethnicity for students enrolled at the university in Fall 2008.

a. Use the information in the graphical display to construct a single segmented bar graph for the ethnicity data.

b. Do you think that the original graphical display or the one you created in Part (a) is more informative? Explain your choice.

c. Why do you think that the original graphical display format (combination of pie chart and segmented bar graph) was chosen over a single pie chart with 7 slices?

3.48 The accompanying graph appeared in *USA Today* (August 5, 2008). This graph is a modified comparative bar graph. Most likely, the modifications (incorporating hands and the earth) were made to try to make a display that readers would find more interesting.

a. Use the information in the *USA Today* graph to construct a traditional comparative bar graph.

b. Explain why the modifications made in the *USA Today* graph may make interpretation more difficult than with the traditional comparative bar graph.

Image not available due to copyright restrictions

Bold exercises answered in back ● Data set available online ✦ Video Solution available

3.49 The two graphical displays below appeared in *USA Today* (June 8, 2009 and July 28, 2009). One is an appropriate representation and the other is not. For each of the two, explain why it is or is not drawn appropriately.

Image not available due to copyright restrictions

Image not available due to copyright restrictions

3.50 The following graphical display is meant to be a comparative bar graph (*USA Today*, August 3, 2009). Do you think that this graphical display is an effective summary of the data? If so, explain why. If not, explain why not and construct a display that makes it easier to compare the ice cream preferences of men and women.

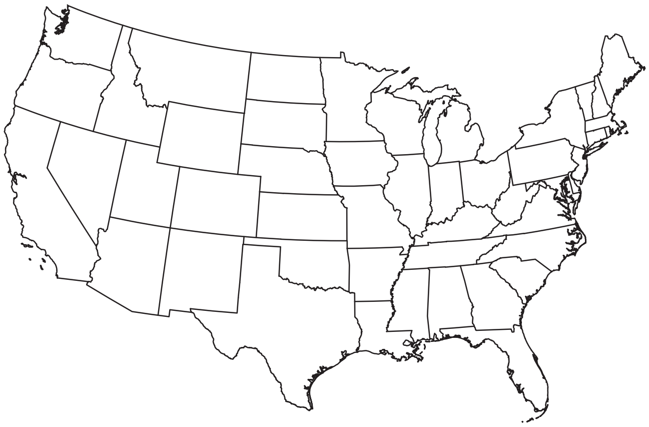
Image not available due to copyright restrictions

Image not available due to copyright restrictions

ACTIVITY 3.1 Locating States

Background: A newspaper article bemoaning the state of students' knowledge of geography claimed that more students could identify the island where the 2002 season of the TV show *Survivor* was filmed than could locate Vermont on a map of the United States. In this activity, you will collect data that will allow you to estimate the proportion of students who can correctly locate the states of Vermont and Nebraska.

1. Working as a class, decide how you will select a sample that you think will be representative of the students from your school.
2. Use the sampling method from Step 1 to obtain the subjects for this study. Subjects should be shown the accompanying map of the United States and asked to point out the state of Vermont. After the subject has given his or her answer, ask the subject to point out the state of Nebraska. For each subject, record whether or not Vermont was correctly identified and whether or not Nebraska was correctly identified.



3. When the data collection process is complete, summarize the resulting data in a table like the one shown here:

Response	Frequency
Correctly identified both states	
Correctly identified Vermont but not Nebraska	
Correctly identified Nebraska but not Vermont	
Did not correctly identify either state	

4. Construct a pie chart that summarizes the data in the table from Step 3.
5. What proportion of sampled students were able to correctly identify Vermont on the map?
6. What proportion of sampled students were able to correctly identify Nebraska on the map?
7. Construct a comparative bar chart that shows the proportion correct and the proportion incorrect for each of the two states considered.
8. Which state, Vermont or Nebraska, is closer to the state in which your school is located? Based on the pie chart, do you think that the students at your school were better able to identify the state that was closer than the one that was farther away? Justify your answer.
9. Write a paragraph commenting on the level of knowledge of U.S. geography demonstrated by the students participating in this study.
10. Would you be comfortable generalizing your conclusions in Step 8 to the population of students at your school? Explain why or why not.

ACTIVITY 3.2 Bean Counters!

Materials needed: A large bowl of dried beans (or marbles, plastic beads, or any other small, fairly regular objects) and a coin.

In this activity, you will investigate whether people can hold more in the right hand or in the left hand.

1. Flip a coin to determine which hand you will measure first. If the coin lands heads side up, start with the right hand. If the coin lands tails side up, start with the left hand. With the designated hand, reach into the bowl and grab as many beans as possible. Raise the hand over the bowl and count to 4. If no beans drop during the count to 4, drop the beans onto a piece of paper and record the number of beans grabbed. If any beans drop during the count, restart the count. That is, you must hold the beans for a count of 4 without any beans falling before you can determine the number grabbed. Repeat the process with the other hand, and then record the following information: (1) right-hand number, (2) left-hand number, and (3) dominant hand (left or right, depending on whether you are left- or right-handed).

2. Create a class data set by recording the values of the three variables listed in Step 1 for each student in your class.
3. Using the class data set, construct a comparative stem-and-leaf display with the right-hand counts displayed on the right and the left-hand counts displayed on the left of the stem-and-leaf display. Comment on the interesting features of the display and include a comparison of the right-hand count and left-hand count distributions.
4. Now construct a comparative stem-and-leaf display that allows you to compare dominant-hand count to nondominant-hand count. Does the display support the theory that dominant-hand count tends to be higher than nondominant-hand count?
5. For each observation in the data set, compute the difference

   dominant-hand count − nondominant-hand count

   Construct a stem-and-leaf display of the differences. Comment on the interesting features of this display.
6. Explain why looking at the distribution of the differences (Step 5) provides more information than the comparative stem-and-leaf display (Step 4). What information is lost in the comparative display that is retained in the display of the differences?

Summary of Key Concepts and Formulas

TERM OR FORMULA	COMMENT
Frequency distribution	A table that displays frequencies, and sometimes relative and cumulative relative frequencies, for categories (categorical data), possible values (discrete numerical data), or class intervals (continuous data).
Comparative bar chart	Two or more bar charts that use the same set of horizontal and vertical axes.
Pie chart	A graph of a frequency distribution for a categorical data set. Each category is represented by a slice of the pie, and the area of the slice is proportional to the corresponding frequency or relative frequency.
Segmented bar graph	A graph of a frequency distribution for a categorical data set. Each category is represented by a segment of the bar, and the area of the segment is proportional to the corresponding frequency or relative frequency.
Stem-and-leaf display	A method of organizing numerical data in which the stem values (leading digit(s) of the observations) are listed in a column, and the leaf (trailing digit(s)) for each observation is then listed beside the corresponding stem. Sometimes stems are repeated to stretch the display.
Histogram	A picture of the information in a frequency distribution for a numerical data set. A rectangle is drawn above each possible value (discrete data) or class interval. The rectangle's area is proportional to the corresponding frequency or relative frequency.
Histogram shapes	A (smoothed) histogram may be unimodal (a single peak), bimodal (two peaks), or multimodal. A unimodal histogram may be symmetric, positively skewed (a long right or upper tail), or negatively skewed. A frequently occurring shape is one that is approximately normal.
Cumulative relative frequency plot	A graph of a cumulative relative frequency distribution.
Scatterplot	A picture of bivariate numerical data in which each observation (x, y) is represented as a point with respect to a horizontal x-axis and a vertical y-axis.
Time-series plot	A graphical display of numerical data collected over time.

Chapter Review Exercises 3.52 - 3.71

3.52 The article "Most Smokers Wish They Could Quit" (*Gallup Poll Analyses*, November 21, 2002) noted that smokers and nonsmokers perceive the risks of smoking differently. The accompanying relative frequency table summarizes responses regarding the perceived harm of smoking for each of three groups: a sample of 241 smokers, a sample of 261 former smokers, and a sample of 502 nonsmokers. Construct a comparative bar chart for these data. Do not forget to use relative frequencies in constructing the bar chart because the three sample sizes are different. Comment on how smokers, former smokers, and nonsmokers differ with respect to perceived risk of smoking.

	Frequency		
Perceived Risk of Smoking	Smokers	Former Smokers	Nonsmokers
Very harmful	145	204	432
Somewhat harmful	72	42	50
Not too harmful	17	10	15
Not at all harmful	7	5	5

3.53 Each year the College Board publishes a profile of students taking the SAT. In the report "2005 College Bound Seniors: Total Group Profile Report," the average SAT scores were reported for three groups defined by first language learned. Use the data in the accompanying table to construct a bar chart of the average verbal SAT score for the three groups.

First Language Learned	Average Verbal SAT
English	519
English and another language	486
A language other than English	462

3.54 The report referenced in Exercise 3.53 also gave average math SAT scores for the three language groups, as shown in the following table.

First Language Learned	Average Math SAT
English	521
English and another language	513
A language other than English	521

Construct a comparative bar chart for the average verbal and math scores for the three language groups. Write a few sentences describing the differences and similarities between the three language groups as shown in the bar chart.

3.55 ● The Connecticut Agricultural Experiment Station conducted a study of the calorie content of different types of beer. The calorie content (calories per 100 ml) for 26 brands of light beer are (from the web site brewery.org):

29	28	33	31	30	33	30	28	27	41	39	31	29
23	32	31	32	19	40	22	34	31	42	35	29	43

Construct a stem-and-leaf display using stems 1, 2, 3, and 4. Write a sentence or two describing the calorie content of light beers.

3.56 The stem-and-leaf display of Exercise 3.16 uses only four stems. Construct a stem-and-leaf display for these data using repeated stems 1H, 2L, 2H, . . . , 4L. For example, the first observation, 29, would have a stem of 2 and a leaf of 9. It would be entered into the display for the stem 2H, because it is a "high" 2—that is, it has a leaf that is on the high end (5, 6, 7, 8, 9).

3.57 ● The article "A Nation Ablaze with Change" (*USA Today*, July 3, 2001) gave the accompanying data on percentage increase in population between 1990 and 2000 for the 50 U.S. states. Also provided in the table is a column that indicates for each state whether the state is in the eastern or western part of the United States (the states are listed in order of population size):

State	Percentage Change	East/West
California	13.8	W
Texas	22.8	W
New York	5.5	E
Florida	23.5	E
Illinois	8.6	E
Pennsylvania	3.4	E
Ohio	4.7	E
Michigan	6.9	E
New Jersey	8.9	E
Georgia	26.4	E
North Carolina	21.4	E

(continued)

Bold exercises answered in back ● Data set available online ✦ Video Solution available

State	Percentage Change	East/West
Virginia	14.4	E
Massachusetts	5.5	E
Indiana	9.7	E
Washington	21.1	W
Tennessee	16.7	E
Missouri	9.3	E
Wisconsin	9.6	E
Maryland	10.8	E
Arizona	40.0	W
Minnesota	12.4	E
Louisiana	5.9	E
Alabama	10.1	E
Colorado	30.6	W
Kentucky	9.7	E
South Carolina	15.1	E
Oklahoma	9.7	W
Oregon	20.4	W
Connecticut	3.6	E
Iowa	5.4	E
Mississippi	10.5	E
Kansas	8.5	W
Arkansas	13.7	E
Utah	29.6	W
Nevada	66.3	W
New Mexico	20.1	W
West Virginia	0.8	E
Nebraska	8.4	W
Idaho	28.5	W
Maine	3.9	E
New Hampshire	11.4	E
Hawaii	9.3	W
Rhode Island	4.5	E
Montana	12.9	W
Delaware	17.6	E
South Dakota	8.5	W
North Dakota	0.5	W
Alaska	14.0	W
Vermont	8.2	E
Wyoming	8.9	W

a. Construct a stem-and-leaf display for percentage growth for the data set consisting of all 50 states. Hints: Regard the observations as having two digits to the left of the decimal place. That is, think of an observation such as 8.5 as 08.5. It will also be easier to truncate leaves to a single digit; for example, a leaf of 8.5 could be truncated to 8 for purposes of constructing the display.

b. Comment on any interesting features of the data set. Do any of the observations appear to be outliers?

c. Now construct a comparative stem-and-leaf display for the eastern and western states. Write a few sentences comparing the percentage growth distributions for eastern and western states.

3.58 ● People suffering from Alzheimer's disease often have difficulty performing basic activities of daily living (ADLs). In one study ("Functional Status and Clinical Findings in Patients with Alzheimer's Disease," *Journal of Gerontology* [1992]: 177–182), investigators focused on six such activities: dressing, bathing, transferring, toileting, walking, and eating. Here are data on the number of ADL impairments for each of 240 patients:

Number of impairments	0	1	2	3	4	5	6
Frequency	100	43	36	17	24	9	11

a. Determine the relative frequencies that correspond to the given frequencies.

b. What proportion of these patients had at most two impairments?

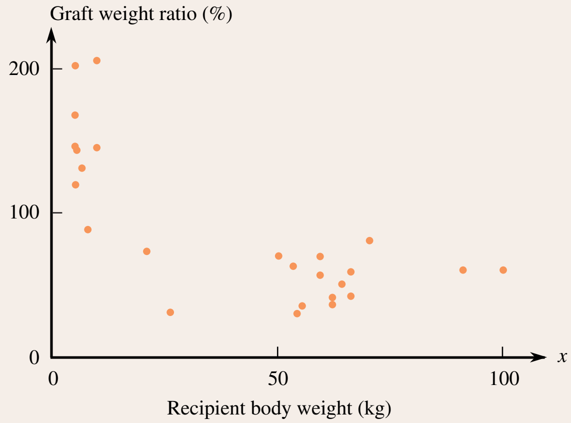
c. Use the result of Part (b) to determine what proportion of patients had more than two impairments.

d. What proportion of the patients had at least four impairments?

3.59 Does the size of a transplanted organ matter? A study that attempted to answer this question ("Minimum Graft Size for Successful Living Donor Liver Transplantation," *Transplantation* [1999]:1112–1116) presented a scatterplot much like the following ("graft weight ratio" is the weight of the transplanted liver relative to the ideal size liver for the recipient):

a. Discuss interesting features of this scatterplot.

b. Why do you think the overall relationship is negative?



3.60 ● The National Telecommunications and Information Administration published a report titled "Falling Through the Net: Toward Digital Inclusion" (U.S. Department of Commerce, October 2000) that included the following information on access to computers in the home:

Year	Percentage of Households with a Computer
1985	8.2
1990	15.0
1994	22.8
1995	24.1
1998	36.6
1999	42.1
2000	51.0

a. Construct a time-series plot for these data. Be careful—the observations are not equally spaced in time. The points in the plot should not be equally spaced along the *x*-axis.

b. Comment on any trend over time.

3.61 According to the National Association of Home Builders, the average size of a home in 1950 was 983 ft^2. The average size increased to 1500 ft^2 in 1970, 2080 ft^2 in 1990; and 2330 ft^2 in 2003 (*San Luis Obispo Tribune*, October 16, 2005).

a. Construct a time-series plot that shows how the average size of a home has changed over time.

b. If the trend of the time-series plot were to continue, what would you predict the average home size to be in 2010?

3.62 The paper "Community Colleges Start to Ask, Where Are the Men?" (*Chronicle of Higher Education*, June 28, 2002) gave data on gender for community college students. It was reported that 42% of students enrolled at community colleges nationwide were male and 58% were female. Construct a segmented bar graph for these data.

3.63 ● The article "Tobacco and Alcohol Use in G-Rated Children's Animated Films" (*Journal of the American Medical Association* [1999]: 1131–1136) reported exposure to tobacco and alcohol use in all G-rated animated films released between 1937 and 1997 by five major film studios. The researchers found that tobacco use was shown in 56% of the reviewed films. Data on the total tobacco exposure time (in seconds) for films with tobacco use produced by Walt Disney, Inc., were as follows:

223 176 548 37 158 51 299 37 11 165
74 92 6 23 206 9

Data for 11 G-rated animated films showing tobacco use that were produced by MGM/United Artists, Warner Brothers, Universal, and Twentieth Century Fox were also given. The tobacco exposure times (in seconds) for these films was as follows:

205 162 6 1 117 5 91 155 24 55 17

Construct a comparative stem-and-leaf display for these data. Comment on the interesting features of this display.

3.64 ● The accompanying data on household expenditures on transportation for the United Kingdom appeared in "Transport Statistics for Great Britain: 2002 Edition" (in *Family Spending: A Report on the Family Expenditure Survey* [The Stationary Office, 2002]). Expenditures (in pounds per week) included costs of purchasing and maintaining any vehicles owned by members of the household and any costs associated with public transportation and leisure travel.

Year	Average Transportation	Percentage of Household Expenditures for Transportation
1990	247.20	16.2
1991	259.00	15.3
1992	271.80	15.8
1993	276.70	15.6
1994	283.60	15.1
1995	289.90	14.9
1996	309.10	15.7
1997	328.80	16.7
1998	352.20	17.0
1999	359.40	17.2
2000	385.70	16.7

a. Construct time-series plots of the transportation expense data and the percent of household expense data.

b. Do the time-series plots of Part (a) support the statement that follows? Explain why or why not. Statement: Although actual expenditures have been increasing, the percentage of the total household expenditures that go toward transportation has remained relatively stable.

Bold exercises answered in back ● Data set available online ✦ Video Solution available

3.65 The article "The Healthy Kids Survey: A Look at the Findings" (*San Luis Obispo Tribune,* October 25, 2002) gave the accompanying information for a sample of fifth graders in San Luis Obispo County. Responses are to the question:

> "After school, are you home alone without adult supervision?"

Response	Percentage
Never	8
Some of the time	15
Most of the time	16
All of the time	61

a. Summarize these data using a pie chart.
b. Construct a segmented bar graph for these data.
c. Which graphing method—the pie chart or the segmented bar graph—do you think does a better job of conveying information about response? Explain.

3.66 "If you were taking a new job and had your choice of a boss, would you prefer to work for a man or a woman?" That was the question posed to individuals in a sample of 576 employed adults (*Gallup at a Glance,* October 16, 2002). Responses are summarized in the following table:

Response	Frequency
Prefer to work for a man	190
Prefer to work for a woman	92
No difference	282
No opinion	12

a. Construct a pie chart to summarize this data set, and write a sentence or two summarizing how people responded to this question.
b. Summarize the given data using a segmented bar graph.

3.67 ● 2005 was a record year for hurricane devastation in the United States (*San Luis Obispo Tribune,* November 30, 2005). Of the 26 tropical storms and hurricanes in the season, four hurricanes hit the mainland: Dennis, Katrina, Rita, and Wilma. The United States insured catastrophic losses since 1989 (approximate values read from a graph that appeared in the *San Luis Obispo Tribune,* November 30, 2005) are as follows:

Year	Cost (in billions of dollars)
1989	7.5
1990	2.5
1991	4.0
1992	22.5
1993	5.0
1994	18.0
1995	9.0
1996	8.0
1997	2.6
1998	10.0
1999	9.0
2000	3.0
2001	27.0
2002	5.0
2003	12.0
2004	28.5
2005	56.8

Construct a time-series plot that shows the insured catastrophic loss over time. What do you think causes the peaks in the graph?

3.68 An article in the *San Luis Obispo Tribune* (November 20, 2002) stated that 39% of those with critical housing needs (those who pay more than half their income for housing) lived in urban areas, 42% lived in suburban areas, and the rest lived in rural areas. Construct a pie chart that shows the distribution of type of residential area (urban, suburban, or rural) for those with critical housing needs.

3.69 ● Living-donor kidney transplants are becoming more common. Often a living donor has chosen to donate a kidney to a relative with kidney disease. The following data appeared in a *USA Today* article on organ transplants ("Kindness Motivates Newest Kidney Donors," June 19, 2002):

	Number of Kidney Transplants	
Year	Living-Donor to Relative	Living-Donor to Unrelated Person
1994	2390	202
1995	2906	400
1996	2916	526
1997	3144	607
1998	3324	814
1999	3359	930
2000	3679	1325
2001	3879	1399

a. Construct a time-series plot for the number of living-donor kidney transplants where the donor is a relative of the recipient. Describe the trend in this plot.

b. Use the data from 1994 and 2001 to construct a comparative bar chart for the type of donation (relative or unrelated). Write a few sentences commenting on your display.

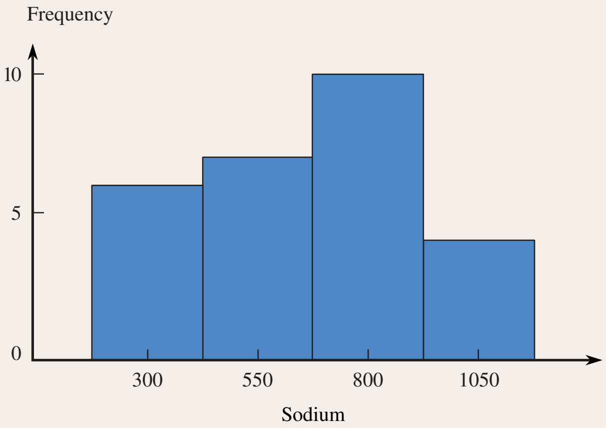
3.70 ● Many nutritional experts have expressed concern about the high levels of sodium in prepared foods. The following data on sodium content (in milligrams) per frozen meal appeared in the article "Comparison of 'Light' Frozen Meals" (Boston Globe, April 24, 1991):

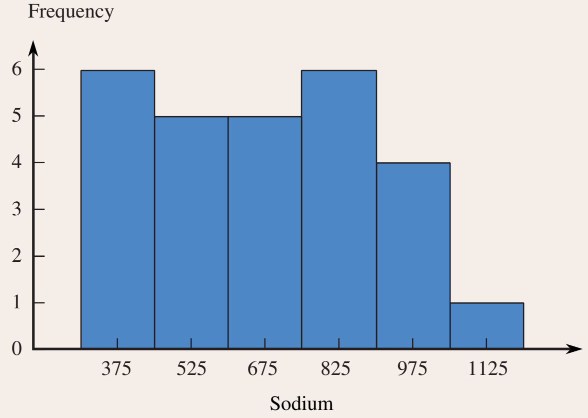
720	530	800	690	880	1050	340	810	760
300	400	680	780	390	950	520	500	630
480	940	450	990	910	420	850	390	600

Two histograms for these data are shown:

a. Do the two histograms give different impressions about the distribution of values?

b. Use each histogram to determine approximately the proportion of observations that are less than 800, and compare to the actual proportion.





3.71 ● Americium 241 (^{241}Am) is a radioactive material used in the manufacture of smoke detectors. The article "Retention and Dosimetry of Injected ^{241}Am in Beagles" (*Radiation Research* [1984]: 564–575) described a study in which 55 beagles were injected with a dose of ^{241}Am (proportional to each animal's weight). Skeletal retention of ^{241}Am (in microcuries per kilogram) was recorded for each beagle, resulting in the following data:

0.196	0.451	0.498	0.411	0.324	0.190	0.489
0.300	0.346	0.448	0.188	0.399	0.305	0.304
0.287	0.243	0.334	0.299	0.292	0.419	0.236
0.315	0.447	0.585	0.291	0.186	0.393	0.419
0.335	0.332	0.292	0.375	0.349	0.324	0.301
0.333	0.408	0.399	0.303	0.318	0.468	0.441
0.306	0.367	0.345	0.428	0.345	0.412	0.337
0.353	0.357	0.320	0.354	0.361	0.329	

a. Construct a frequency distribution for these data, and draw the corresponding histogram.

b. Write a short description of the important features of the shape of the histogram.

Graphing Calculator Explorations

EXPLORATION 3.1 Using Lists on Your Calculator

Calculators and computers work their magic by storing numbers in "memory locations." To perform an addition, the computer looks in its "memory" for the two numbers, retrieves them, and adds them. In the early days of calculators there were very few of these expensive memory cells and calculations were performed one at a time while the user entered numbers. The modern scientific calculator allows a very useful extension of a single memory cell: a *group* of memory cells known as a "list." Using a list, a whole set of data, complete with a "name," can be stored in the calculator and analyzed as a whole. This list capability makes the calculator a very powerful tool for analyzing data. Since all the numbers are in the calculator at the same time, graphic representations of data such as those presented in this chapter are possible.

The actual capabilities of lists and the keystrokes to use these capabilities vary from calculator to calculator, so we will not be overly specific about particular calculator keystrokes. *Your best source for learning about your calculator is the manual that came with it!* You need the manual to fully understand and realize the potential of your calculator. Your calculator may implement some of its capabilities with special keys or menus, and the menus may include functions that require additional information that must be entered in a particular order. You don't need to memorize all these details. That's why you have the manual!

To use your calculator for statistical analysis you must be able to manipulate lists effectively. The accompanying table describes some of the list-based calculator features that will be important to you in performing statistical analysis. In future Explorations, we will assume that you are familiar with these list features.

Capability	Why It Is Important
Create a list	Some lists are provided automatically for your use, already labeled "List 1 or L1," or "List 2 or L2," etc. You will want to save data in lists with more informative names than these, such as "height" or "time."
Enter data into a list	This knowledge is, of course, fundamental to all analyses you will be performing.
Delete a list from the calculator	As powerful as your calculator is, there is a limit to its memory—there will come a time when you will need to delete the old data to prepare for the new.
Insert a number in a certain location in the list Or Delete a particular number in the list	If you are like everyone else, you will eventually add an extra number you did not intend or leave out a number from where it should be. Correcting these errors is a lot faster than deleting a whole list and starting over.
Copy data from one list to another	This will give you one of those things so precious to everyone who works with a finicky calculator (or finicky fingers?): a backup copy of the data!
Perform arithmetic operations with lists	Rather than perform the same arithmetic sequence separately on a set of numbers, you can do the calculations one list at a time. For example, to change units from inches to centimeters, you can multiply *all* the numbers on the list by 2.54. Usually, this is done with a statement something like 2.54 × ListName1 → ListName2. (The equal sign, =, is sometimes used in place of the arrow symbol.)

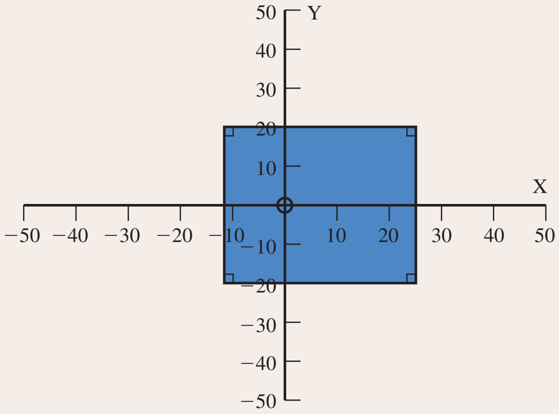
You need to be familiar with list manipulation to do effective statistical work with your calculator. As we proceed we will be more specific—and more detailed—

about how you can utilize your calculator's list capabilities. The calculator manual may have a chapter called "Using Lists." We strongly encourage you to read it!

EXPLORATION 3.2 Setting the Statistics Window

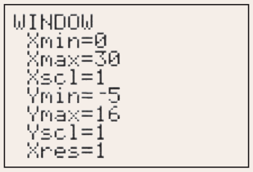
As we move into graphical descriptions of data, we need to consider how to use your calculator to produce graphs and plots of data. If you have used your calculator to graph functions, some of what follows will be a review. If you are new to the world of graphing calculators you will need a basic understanding of how to set up your calculator's "viewing window" for displaying graphs.

The metaphor of viewing the "world" through a "window" is a good one for thinking about the calculator window. If you think of the Cartesian *x*-*y* axes as the calculator's "world view" and your calculator view window as a portal through which to view this Cartesian world, you will be in the right frame of mind.



As you set up your calculator for graphing, your first problem will be, "Where in the *world* do I put my view *window?*" The quick and easy answer: You put your view window where your data is! We will illustrate how to do this by constructing a histogram, using the data from Exercise 3.17, "Going wireless." To illustrate how to set the view window, we will begin with a slightly bad graph of a histogram and then gradually improve it. First, we want you to do something a bit strange, but trust us. Enter the function $y = 100/x$ into your calculator. Now enter the data from "Going wireless" into your calculator in List 1. After entering the data, navigate your calculator's menu system to the histogram option. Actual keystrokes will vary among calculators, but the terms "Stat Graph" and "Stat Plot" are commonly used in calculators.

The properties of the view window are based on settings that you will manually enter into the calculator. (Your calculator may have the capability of automatically placing the view window over the appropriate position in its Cartesian system. Pretend for the moment that you are unaware of this.) To set up the view window you must navigate to your calculator's menu system, or possibly just press a "window" key. When you find the graph setup screen, it will look something like Figure 3.42. (There may be different or additional information on your screen but the numbers here will be our focus for this Exploration. Your numbers may have different values from these; for ease in following the discussion you may wish to change the values on your calculator to match those in the figure.)



```
WINDOW
 Xmin=0
 Xmax=30
 Xscl=1
 Ymin=-5
 Ymax=16
 Yscl=1
 Xres=1
```

FIGURE 3.42

The numbers on this screen determine where the viewing window is placed over the calculator's "world" coordinate system. Exit from the setup window and plot a histogram, by using a "graph" key or a menu, depending on your calculator. You should have something like what is shown in Figure 3.43. (Your graph will get better as we go.)

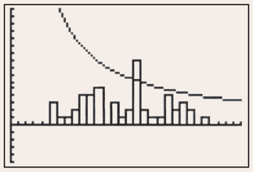


FIGURE 3.43

Notice that there actually is a histogram there, but there is a pesky function overlaying the histogram. Suspiciously, it looks something like $y = 100/x$! What's the problem here? First, remember that the calculator's world is the x-y coordinate system. If you have been using the calculator to plot mathematical functions and they are still "in" the calculator, both your last function and the histogram will be drawn. Oddly, this happens by design. Since a calculator user most likely will not graph a mathematical function and a histogram at the same time, the calculator graphing space is shared to save calculator memory. The solution to the histogram-and-function problem is easy: Don't graph the function. You can now delete the function and redraw the histogram.

We will need to manually reposition the viewing window in the calculator world for a better look at the data. Return to the view window setup screen and make the following changes: set Xmin to 4 and Xmax to 28 and redraw the histogram. The histogram in the view window will now be positioned something like Figure 3.44.

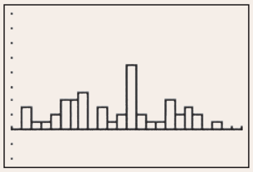


FIGURE 3.44

This is certainly an improvement but the histogram is still rather small compared to the screen, sacrificing some detail. To correct this problem we will adjust the top of the view window. Return to the graph setup screen and locate the lines for YMin and YMax. Change the YMin to -1 and the YMax to 16. Now regraph the histogram. You should see a very well-spaced histogram similar to Figure 3.45.

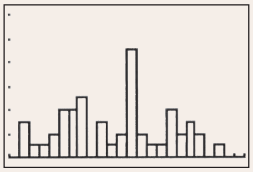
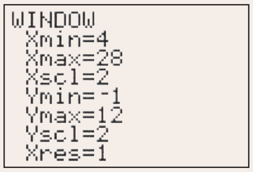


FIGURE 3.45

You will find that adjusting the view window is a frequent task in creating effective statistical graphs. While each of the statistical plots has its own individuality, your construction of them will always involve positioning the view window over the Cartesian world view of your calculator. Even if the calculator has an automatic function to position statistical graphs, you will find it necessary sometimes to "improve" on the calculator's automatic choice. When you manually change your graph, keep in mind the idea of positioning a view window over a Cartesian coordinate system. This will help you organize your thoughts about how to change the view window and make the task less frustrating.

EXPLORATION 3.3 Scaling the Histogram

When we constructed a histogram in the previous Exploration there were some numbers that we temporarily ignored in the view screen. We would like to return to those numbers now because they can seriously affect the look of a histogram. When we left the histogram the numbers in our view window were set as shown in Figure 3.46. These settings place the view window over the calculator's Cartesian system for effective viewing of the histogram from the "Going wireless" data.


```
WINDOW
 Xmin=4
 Xmax=28
 Xscl=2
 Ymin=-1
 Ymax=12
 Yscl=2
 Xres=1
```
FIGURE 3.46

We would now like to experiment a bit with the "Xscale." In the statistical graphs produced by the calculator the Xscale and Yscale choices will control the placement of the little "tick" marks on the x- and y-axis. In Exploration 3.2, the Xscale and Yscale were set at 1. Change the Xscale and Yscale values to 2 and redraw the histogram. You should see a graph similar to Figure 3.47. The x- and y-axis tick marks now appear at multiples of 2.

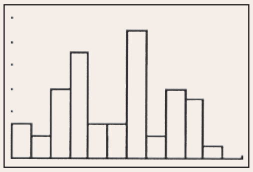


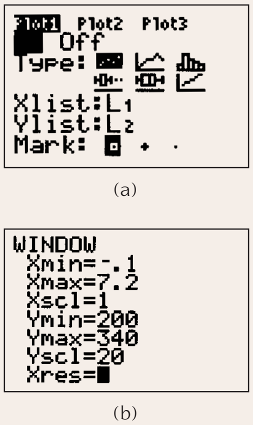
FIGURE 3.47

Note that changing the Xscale has altered not only the tick marks but also the class intervals for the histogram. The choice of class intervals can significantly change the look and feel of the histogram. The choice of Xscale can affect judgments about the shape of the histogram. Because of this possibility it is wise to look at a histogram with varying choices of the Xscale value. If the shape appears very similar for different choices of Xscale, you can interpret and describe the shape with more confidence. However, if different Xscale choices alter the look of the histogram you should probably be more tentative.

EXPLORATION 3.4 The Scatterplot

In this Exploration, we consider graphing a scatterplot of bivariate data. Here are the steps for creating a scatterplot:

1. Navigate your calculator's menu system to select the type of graph you want.
2. Select an appropriate viewing window.
3. Select data from two lists rather than one.
4. Indicate which list will correspond to the horizontal axis and which will correspond to the vertical axis.



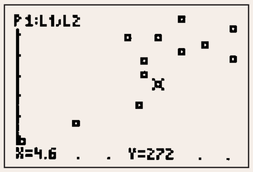
```
Plot1 Plot2 Plot3
On Off
Type:
Xlist:L1
Ylist:L2
Mark: □ + .
```

(a)

```
WINDOW
 Xmin=-.1
 Xmax=7.2
 Xscl=1
 Ymin=200
 Ymax=340
 Yscl=20
 Xres=█
```

(b)

FIGURE 3.48



```
P1:L1,L2
X=4.6         Y=272
```

FIGURE 3.49

Figure 3.48(a) shows the selection for a scatterplot. We have selected graphing parameters and scale information as indicated Figure 3.48(b). The resulting plot is shown in Figure 3.49. The *x* and *y* scales are set so that the points will fill the view screen. Notice also that the plot is set to be made up of small squares—they are more easily seen than small dots.

One problem that frequently arises with scatterplots on calculators is the "disappearing" axis (in this case, the horizontal axis) and the lack of a discernible scale on either axis. The calculator has been hardwired to plot the *x*-axis and *y*-axis, but these will not be displayed if the view screen is not set up for both positive and negative values for the axes. Unfortunately the demands for axes as reference points compete with the desire for detail in a scatter plot. Although selecting a *y* scale from −1 to 200 would show each axis, the points would cluster close to the top of the screen, possibly hiding some detail or pattern from the data analyst.

There are no easy solutions to this problem, but use of the "trace" capability on your calculator can help a little. Pressing the trace button causes the calculator to display the coordinates of the location of a little crosshair icon. You can move this icon about the screen by pressing special "arrow" keys on your calculator. If you are displaying data rather than a function, pressing the trace button and then arrow keys may move the crosshair icon from point to point, displaying the coordinates of the points of your scatter plot. The scatter plot in Figure 3.49 shows the crosshair icon on a point (4.6, 272).

In your classroom your instructor may want you to sketch a scatter plot on a particular assignment, and most likely will want some sort of axes and scales. The scaling information is easily found in the viewing window and translated to the sketch of the plot on your paper.

Cumulative Review Exercises CR3.1 - CR3.16

CR3.1 Does eating broccoli reduce the risk of prostate cancer? According to an observational study from the Fred Hutchinson Cancer Research Center (see the CNN.com web site article titled "Broccoli, Not Pizza Sauce, Cuts Cancer Risk, Study Finds," January 5, 2000), men who ate more cruciferous vegetables (broccoli, cauliflower, brussels sprouts, and cabbage) had a lower risk of prostate cancer. This study made separate comparisons for men who ate different levels of vegetables. According to one of the investigators, "at any given level of total vegetable consumption, as the percent of cruciferous vegetables increased, the prostate cancer risk decreased." Based on this study, is it reasonable to conclude that eating cruciferous vegetables causes a reduction in prostate cancer risk? Explain.

CR3.2 An article that appeared in *USA Today* (August 11, 1998) described a study on prayer and blood pressure. In this study, 2391 people 65 years or older, were followed for 6 years. The article stated that people who attended a religious service once a week and prayed or studied the Bible at least once a day were less likely to

Bold exercises answered in back ● Data set available online ✦ Video Solution available

have high blood pressure. The researcher then concluded that "attending religious services lowers blood pressure". The headline for this article was "Prayer Can Lower Blood Pressure." Write a few sentences commenting on the appropriateness of the researcher's conclusion and on the article headline.

CR3.3 Sometimes samples are composed entirely of volunteer responders. Give a brief description of the dangers of using voluntary response samples.

CR3.4 A newspaper headline stated that at a recent budget workshop, nearly three dozen people supported a sales tax increase to help deal with the city's financial deficit (*San Luis Obispo Tribune,* January 22, 2005). This conclusion was based on data from a survey acknowledged to be unscientific, in which 34 out of the 43 people who chose to attend the budget workshop recommended raising the sales tax. Briefly discuss why the survey was described as "unscientific" and how this might limit the conclusions that can be drawn from the survey data.

CR3.5 "More than half of California's doctors say they are so frustrated with managed care they will quit, retire early, or leave the state within three years." This conclusion from an article titled "Doctors Feeling Pessimistic, Study Finds" (*San Luis Obispo Tribune,* July 15, 2001) was based on a mail survey conducted by the California Medical Association. Surveys were mailed to 19,000 California doctors, and 2000 completed surveys were returned. Describe any concerns you have regarding the conclusion drawn.

CR3.6 Based on observing more than 400 drivers in the Atlanta area, two investigators at Georgia State University concluded that people exiting parking spaces did so more slowly when a driver in another car was waiting for the space than when no one was waiting ("Territorial Defense in Parking Lots: Retaliation Against Waiting Drivers," *Journal of Applied Social Psychology* [1997]: 821–834). Describe how you might design an experiment to determine whether this phenomenon is true for your city. What is the response variable? What are some extraneous variables and how does your design control for them?

CR3.7 An article from the Associated Press (May 14, 2002) led with the headline "Academic Success Lowers Pregnancy Risk." The article described an evaluation of a program that involved about 350 students at 18 Seattle schools in high crime areas. Some students took part in a program beginning in elementary school in which teachers showed children how to control their impulses, recognize the feelings of others, and get what they want without aggressive behavior. Others did not participate in the program. The study concluded that the program was effective because by the time young women in the program reached age 21, the pregnancy rate among them was 38%, compared to 56% for the women in the experiment who did not take part in the program. Explain why this conclusion is valid only if the women in the experiment were randomly assigned to one of the two experimental groups.

CR3.8 Researchers at the University of Pennsylvania suggest that a nasal spray derived from pheromones (chemicals emitted by animals when they are trying to attract a mate) may be beneficial in relieving symptoms of premenstrual syndrome (PMS) (*Los Angeles Times,* January 17, 2003).

a. Describe how you might design an experiment using 100 female volunteers who suffer from PMS to determine whether the nasal spray reduces PMS symptoms.

b. Does your design from Part (a) include a placebo treatment? Why or why not?

c. Does your design from Part (a) involve blinding? Is it single-blind or double-blind? Explain.

CR3.9 Students in California are required to pass an exit exam in order to graduate from high school. The pass rate for San Luis Obispo High School has been rising, as have the rates for San Luis Obispo County and the state of California (*San Luis Obispo Tribune,* August 17, 2004). The percentage of students who passed the test was as follows:

Year	District	Pass Rate
2002	San Luis Obispo High School	66%
2003		72%
2004		93%
2002	San Luis Obispo County	62%
2003		57%
2004		85%
2002	State of California	32%
2003		43%
2004		74%

a. Construct a comparative bar chart that allows the change in the pass rate for each group to be compared.

b. Is the change the same for each group? Comment on any difference observed.

CR3.10 A poll conducted by the Associated Press–Ipsos on public attitudes found that most Americans are convinced that political corruption is a major problem (*San Luis Obispo Tribune*, December 9, 2005). In the poll, 1002 adults were surveyed. Two of the questions and the summarized responses to these questions follow:

How widespread do you think corruption is in public service in America?

Hardly anyone	1%
A small number	20%
A moderate number	39%
A lot of people	28%
Almost everyone	10%
Not sure	2%

In general, which elected officials would you say are more ethical?

Democrats	36%
Republicans	33%
Both equally	10%
Neither	15%
Not sure	6%

a. For each question, construct a pie chart summarizing the data.
b. For each question, construct a segmented bar chart displaying the data.
c. Which type of graph (pie chart or segmented bar graph) does a better job of presenting the data? Explain.

CR3.11 ● The article "Determination of Most Representative Subdivision" (*Journal of Energy Engineering* [1993]: 43–55) gave data on various characteristics of subdivisions that could be used in deciding whether to provide electrical power using overhead lines or underground lines. Data on the variable x = total length of streets within a subdivision are as follows:

1280	5320	4390	2100	1240	3060	4770	1050
360	3330	3380	340	1000	960	1320	530
3350	540	3870	1250	2400	960	1120	2120
450	2250	2320	2400	3150	5700	5220	500
1850	2460	5850	2700	2730	1670	100	5770
3150	1890	510	240	396	1419	2109	5770

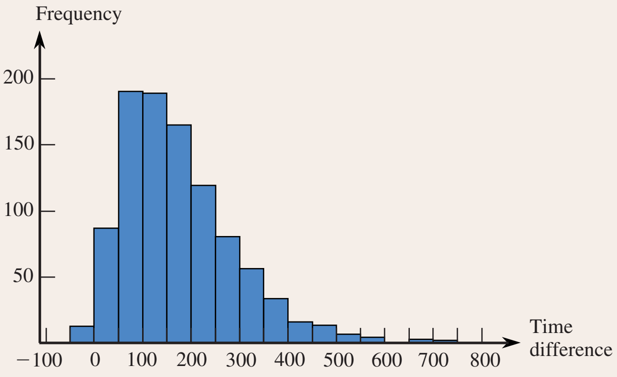
a. Construct a stem-and-leaf display for these data using the thousands digit as the stem. Comment on the various features of the display.
b. Construct a histogram using class boundaries of 0 to <1000, 1000 to <2000, and so on. How would you describe the shape of the histogram?
c. What proportion of subdivisions has total length less than 2000? between 2000 and 4000?

CR3.12 ● The paper "Lessons from Pacemaker Implantations" (*Journal of the American Medical Association* [1965]: 231–232) gave the results of a study that followed 89 heart patients who had received electronic pacemakers. The time (in months) to the first electrical malfunction of the pacemaker was recorded:

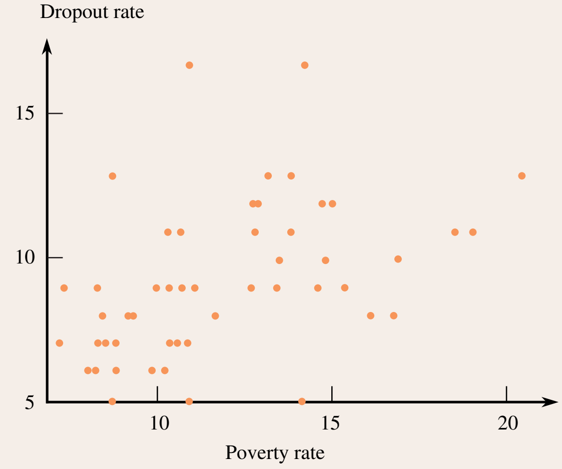
24	20	16	32	14	22	2	12	24	6	10	20
8	16	12	24	14	20	18	14	16	18	20	22
24	26	28	18	14	10	12	24	6	12	18	16
34	18	20	22	24	26	18	2	18	12	12	8
24	10	14	16	22	24	22	20	24	28	20	22
26	20	6	14	16	18	24	18	16	6	16	10
14	18	24	22	28	24	30	34	26	24	22	28
30	22	24	22	328							

a. Summarize these data in the form of a frequency distribution, using class intervals of 0 to <6, 6 to <12, and so on.
b. Compute the relative frequencies and cumulative relative frequencies for each class interval of the frequency distribution of Part (a).
c. Show how the relative frequency for the class interval 12 to <18 could be obtained from the cumulative relative frequencies.
d. Use the cumulative relative frequencies to give approximate answers to the following:
   i. What proportion of those who participated in the study had pacemakers that did not malfunction within the first year?
   ii. If the pacemaker must be replaced as soon as the first electrical malfunction occurs, approximately what proportion required replacement between 1 and 2 years after implantation?
e. Construct a cumulative relative frequency plot, and use it to answer the following questions.
   i. What is the approximate time at which about 50% of the pacemakers had failed?
   ii. What is the approximate time at which only about 10% of the pacemakers initially implanted were still functioning?

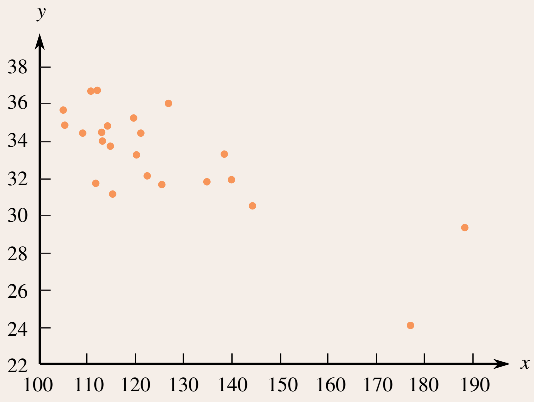
CR3.13 How does the speed of a runner vary over the course of a marathon (a distance of 42.195 km)? Consider determining both the time (in seconds) to run the first 5 km and the time (in seconds) to run between the 35 km and 40 km points, and then subtracting the 5-km time from the 35–40-km time. A positive value of this difference corresponds to a runner slowing down toward the end of the race. The histogram below is based on times of runners who participated in several different Japanese marathons ("Factors Affecting Runners' Marathon Performance," *Chance* [Fall 1993]: 24–30). What are some interesting features of this histogram? What is a typical difference value? Roughly what proportion of the runners ran the late distance more quickly than the early distance?



CR3.14 Data on x = poverty rate (%) and y = high school dropout rate (%) for the 50 U.S. states and the District of Columbia were used to construct the following scatterplot (*Chronicle of Higher Education,* August 31, 2001): Write a few sentences commenting on this scatterplot. Would you describe the relationship between poverty rate and dropout rate as positive (y tends to increase as x increases), negative (y tends to decrease as x increases), or as having no discernible relationship between x and y?



CR3.15 ✦ One factor in the development of tennis elbow, a malady that strikes fear into the hearts of all serious players of that sport, is the impact-induced vibration of the racket-and-arm system at ball contact. It is well known that the likelihood of getting tennis elbow depends on various properties of the racket used. Consider the accompanying scatterplot of x = racket resonance frequency (in hertz) and y = sum of peak-to-peak accelerations (a characteristic of arm vibration, in meters per second per second) for n = 23 different rackets ("Transfer of Tennis Racket Vibrations into the Human Forearm," *Medicine and Science in Sports and Exercise* [1992]: 1134–1140). Discuss interesting features of the data and of the scatterplot.

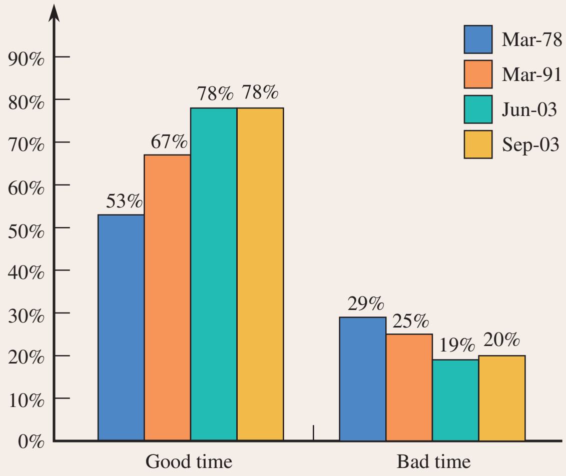


CR3.16 An article that appeared in *USA Today* (September 3, 2003) included a graph similar to the one shown here summarizing responses from polls conducted in 1978, 1991, and 2003 in which a sample of American adults were asked whether or not it was a good time or a bad time to buy a house.

a. Construct a time-series plot that shows how the percentage that thought it was a good time to buy a house has changed over time.

b. Add a new line to the plot from Part (a) showing the percentage that thought it was a bad time to buy a house over time. Be sure to label the lines clearly.

c. Which graph, the given bar chart or the time-series plot, best shows the trend over time?



Bold exercises answered in back ● Data set available online ✦ Video Solution available