2

Organizing Data

In dwelling upon the vital importance of sound observation, it must never be lost sight of what observation is for. It is not for the sake of piling up miscellaneous information or curious facts, but for the sake of saving life and increasing health and comfort.

—Florence Nightingale, *Notes on Nursing*

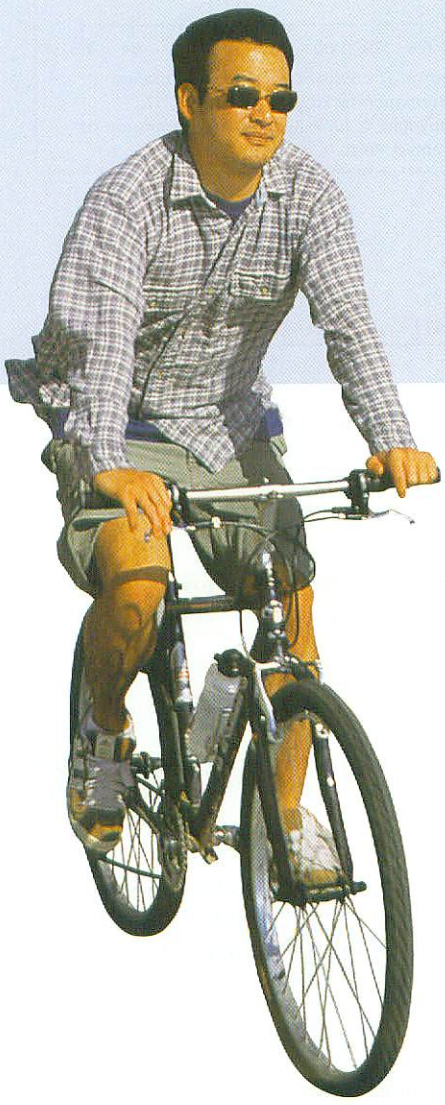
Florence Nightingale (1820–1910)

Florence Nightingale has been described as a "passionate statistician" and a "relevant statistician." She viewed statistics as a science that allowed one to transcend his or her narrow individual experience and aspire to the broader service of humanity. She was one of the first nurses to use graphic representation of statistics, illustrating with charts and diagrams how improved sanitation decreased the rate of mortality. Her statistical reports about the appalling sanitary conditions at Scutari (the main British hospital during the Crimean War) were taken very seriously by the English Secretary at War, Sidney Herbert. When sanitary reforms recommended by Nightingale were instituted in military hospitals, the mortality rate dropped from an incredible 42.7% to only 2.2%.

PREVIEW QUESTIONS

- What are histograms? When are they used? (SECTION 2.1)
- What are common distribution shapes? (SECTION 2.1)
- How can you select graphs appropriate for given data sets? (SECTION 2.2)
- How can you quickly order data and, at the same time, reveal the distribution shape? (SECTION 2.3)

For on-line student resources, visit **math.college.hmco.com/students** and follow the Statistics links to the Brase/Brase, *Understanding Basic Statistics*, 4th edition web site.



FOCUS PROBLEM

Say It with Pictures

Edward R. Tufte, in his book *The Visual Display of Quantitative Information*, presents a number of guidelines for producing good graphics. According to the criteria, a graphical display should

- show the data;
- induce the viewer to think about the substance of the graphic rather than about the methodology, the design, the technology, or other production devices;
- avoid distorting what the data have to say.

As an example of a graph that violates some of the criteria, Tufte includes a graphic that appeared in a well-known newspaper. Figure 2-1(a) shows a facsimile of the problem graphic, whereas part (b) of the figure shows a better rendition of the data display.

After completing this chapter, you will be able to answer the following questions.

(a) Look at the graph in Figure 2-1(a). Is it essentially a bar graph? Explain. What are some of the flaws of Figure 2-1(a) as a bar graph?

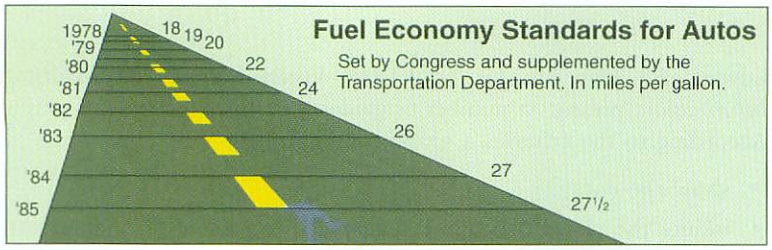
(b) Examine Figure 2-1(b), which shows the same information. Is it essentially a time-series graph? Explain. In what ways does the second graph seem to display the information in a clearer manner?

(See Problem 1 of the Chapter 2 Review Problems.)

FIGURE 2-1

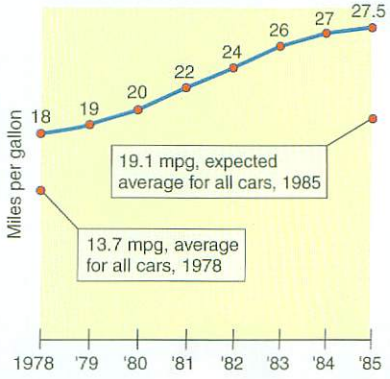
Fuel Economy Standards for Autos

(a)



Source: Copyright © 1978 by The New York Times Company. Reprinted by permission.

(b) Required Fuel Economy Standards: New Cars Built from 1978 to 1985



Source: The Visual Display of Quantitative Information by Edward R. Tufte, p. 57. Copyright © 1983. Reprinted by permission of Graphics Press.

2.1 Frequency Distributions, Histograms, and Related Topics

FOCUS POINTS

- ✓ Organize raw data using a frequency table.
- ✓ Construct histograms and relative-frequency histograms.
- ✓ Recognize basic distribution shapes: uniform, symmetric, skewed, and bimodal.
- ✓ Interpret graphs in the context of the data setting.

Frequency Tables

When we have a large set of quantitative data, it's useful to organize it into smaller intervals or *classes* and count how many data values fall into each class. A frequency table does just that.

A **frequency table** partitions data into classes or intervals and shows how many data values are in each class. The classes or intervals are constructed so that each data value falls into exactly one class.

Constructing a frequency table involves a number of steps. Example 1 demonstrates the steps.

EXAMPLE 1
Frequency table

A task force to encourage car pooling did a study of one-way commuting distances of workers in the downtown Dallas area. A random sample of 60 of these workers was taken. The commuting distances of the workers in the sample are given in Table 2-1. Make a frequency table for these data.

SOLUTION:

(a) First decide how many classes you want. Five to 15 classes are usually used. If you use fewer than five classes, you risk losing too much information. If you use more

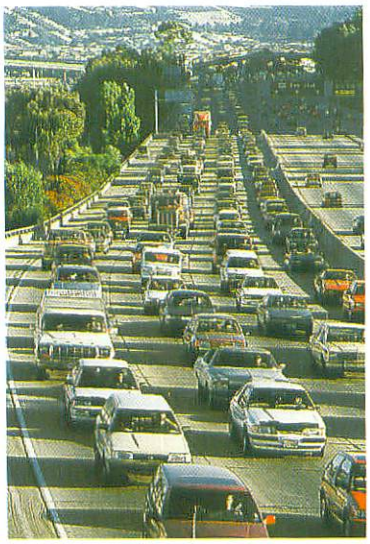


TABLE 2-1 One-Way Commuting Distances (in Miles) for 60 Workers in Downtown Dallas

13	47	10	3	16	20	17	40	4	2
7	25	8	21	19	15	3	17	14	6
12	45	1	8	4	16	11	18	23	12
6	2	14	13	7	15	46	12	9	18
34	13	41	28	36	17	24	27	29	9
14	26	10	24	37	31	8	16	12	16

than 15 classes, the data may not be sufficiently summarized. Let the spread of the data and the purpose of the frequency table be your guide when selecting the number of classes. In the case of the commuting data, let's use *six* classes.

(b) Next, find the *class width* for the six classes.

Class width

PROCEDURE

How to find the class width

1. Compute $\dfrac{\text{Largest data value} - \text{smallest data value}}{\text{Desired number of classes}}$
2. Increase the value computed to the next highest whole number.

Note: To ensure that all the classes taken together cover the data, we need to increase the result of step 1 to the *next whole number,* even if step 1 produced a whole number. For instance, if the calculation in step 1 produces the value 4, we make the class width 5.

To find the class width for the commuting data, we observe that the largest distance commuted is 47 miles and the smallest is 1 mile. Using six classes, the class width is 8, since

$$\text{Class width} = \frac{47 - 1}{6} \approx 7.7 \quad \text{(increase to 8)}$$

(c) Now we determine the data range for each class.

Class limits

The **lower class limit** is the lowest data value that can fit in a class. The **upper class limit** is the highest data value that can fit in a class. The **class width** is the difference between the *lower* class limit of one class and the *lower* class limit of the next class.

The smallest commuting distance in our sample is 1 mile. We use this *smallest* data value as the lower class limit of the *first* class. Since the class width is 8, we add 8 to 1 to find that the lower class limit for the *second* class is 9. Following this pattern, we establish *all* the *lower class limits. Then we fill in the*

TABLE 2-2 Frequency Table of One-Way Commuting Distances for 60 Downtown Dallas Workers (data in miles)

Class Limits Lower–Upper	Class Boundaries Lower–Upper	Tally	Frequency	Class Midpoint
1–8	0.5–8.5	𝍸 𝍸 ||||	14	4.5
9–16	8.5–16.5	𝍸 𝍸 𝍸 𝍸 |	21	12.5
17–24	16.5–24.5	𝍸 𝍸 |	11	20.5
25–32	24.5–32.5	𝍸 |	6	28.5
33–40	32.5–40.5	||||	4	36.5
41–48	40.5–48.5	||||	4	44.5

upper class limits so that the classes span the entire range of data. Table 2-2 shows the upper and lower class limits for the commuting distance data.

(d) Now we are ready to tally the commuting distance data into the six classes and find the frequency for each class.

PROCEDURE

How to tally data

Tallying data is a method of counting data values that fall into a particular class or category.

To tally data into classes of a frequency table, examine each data value. Determine which class contains the data value and make a tally mark or vertical stroke (|) beside that class. For ease of counting, each fifth tally mark of a class is placed diagonally across the prior four marks (𝍸).

Class frequency

The *class frequency* for a class is the number of tally marks corresponding to that class.

Table 2-2 shows the tally and frequency of each class.

Class midpoint or class mark

(e) The center of each class is called the *midpoint* (or *class mark*). The midpoint is often used as a representative value of the entire class. The midpoint is found by adding the lower and upper class limits of one class and dividing by 2.

$$\text{Midpoint} = \frac{\text{Lower class limit} + \text{upper class limit}}{2}$$

Table 2-2 shows the class midpoints.

Class boundaries

(f) There is a space between the upper limit of one class and the lower limit of the next class. The halfway points of these intervals are called *class boundaries*. These are shown in Table 2-2.

PROCEDURE

How to find class boundaries (integer data)

To find **upper class boundaries**, add 0.5 unit to the upper class limits. To find **lower class boundaries**, subtract 0.5 unit from the lower class limits.

Relative frequency

Basic frequency tables show how many data values fall into each class. It's also useful to know the *relative frequency* of a class. The relative frequency of a class is the proportion of all data values that fall into that class. To find the relative frequency of a particular class, divide the class frequency f by the total of all frequencies n (sample size).

$$\text{Relative frequency} = \frac{f}{n} = \frac{\text{Class frequency}}{\text{Total of all frequencies}}$$

Table 2-3 shows the relative frequencies for the commuter data of Table 2-1. Since we already have the frequency table (Table 2-2), the relative-frequency table is obtained easily. The sample size is $n = 60$. Notice that the sample size is the total of all the frequencies. Therefore, the relative frequency for the first class (the class from 1 to 8) is

$$\text{Relative frequency} = \frac{f}{n} = \frac{14}{60} \approx 0.23$$

The symbol $\approx$ means "approximately equal to." We use the symbol because we rounded the relative frequency. Relative frequencies for the other classes are computed in a similar way.

The total of the relative frequencies should be 1. However, rounded results may make the total slightly higher or lower than 1.

TABLE 2-3 Relative Frequencies of One-Way Commuting Distances

Class	Frequency f	Relative Frequency f/n
1–8	14	14/60 ≈ 0.23
9–16	21	21/60 ≈ 0.35
17–24	11	11/60 ≈ 0.18
25–32	6	6/60 ≈ 0.10
33–40	4	4/60 ≈ 0.07
41–48	4	4/60 ≈ 0.07

Let's summarize the procedure for making a frequency table that includes relative frequencies.

PROCEDURE

How to make a frequency table

1. Determine the number of classes and the corresponding class width.
2. Create the distinct classes. We use the convention that the *lower class limit* of the first class is the smallest data value. Add the class width to this number to get the *lower class limit* of the next class.
3. Fill in *upper class limits* to create distinct classes that accommodate all possible data values from the data set.
4. Tally the data into classes. Each data value should fall into exactly one class. Total the tallies to obtain each *class frequency.*
5. Compute the *midpoint* (class mark) for each class.
6. Determine the *class boundaries.*

PROCEDURE

How to make a relative-frequency table

First make a frequency table. Then, for each class, compute the *relative frequency f/n,* where f is the class frequency and n is the total sample size.

Histograms and Relative-Frequency Histograms

Histograms and relative-frequency histograms provide effective visual displays of data organized into frequency tables. In these graphs, we use bars to represent each class, where the width of the bar is the class width. For histograms, the height of the bar is the class frequency, whereas for relative-frequency histograms, the height of the bar is the relative frequency of that class.

PROCEDURE

How to make a histogram or a relative-frequency histogram

1. Make a frequency table (including relative frequencies) with the designated number of classes.
2. Place class boundaries on the horizontal axis and frequencies or relative frequencies on the vertical axis.
3. For each class of the frequency table, draw a bar whose width extends between corresponding class boundaries. For histograms, the height of each bar is the corresponding class frequency. For relative-frequency histograms, the height of each bar is the corresponding class relative frequency.

EXAMPLE 2

Histogram and relative-frequency histogram

Make a histogram and a relative-frequency histogram with six bars for the data in Table 2-1 showing one-way commuting distances.

SOLUTION: The first step is to make a frequency table and a relative-frequency table with six classes. We'll use Table 2-2 and Table 2-3. Figures 2-2 and 2-3 show the histogram and relative-frequency histogram. In both graphs, class boundaries are marked on the horizontal axis. For each class of the frequency table, make a corresponding bar with horizontal width extending from the lower boundary to the upper boundary of the respective class. For a histogram, the height of each bar is the corresponding class frequency. For a relative-frequency histogram, the height of each bar is the corresponding relative frequency. Notice that the basic shapes of the graphs are the same. The only difference involves the vertical axis. The vertical axis of the histogram shows frequencies, whereas that of the relative-frequency histogram shows relative frequencies.

FIGURE 2-2

Histogram for Dallas Commuters: One-Way Commuting Distances

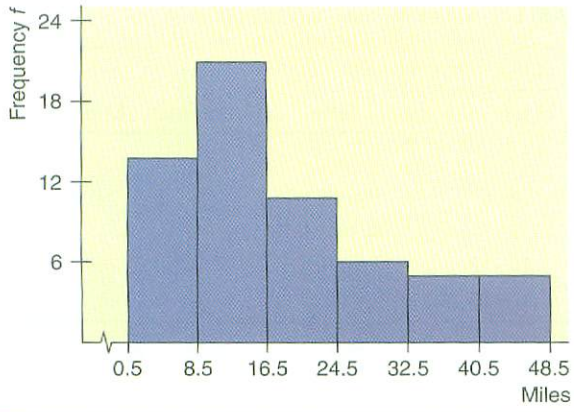
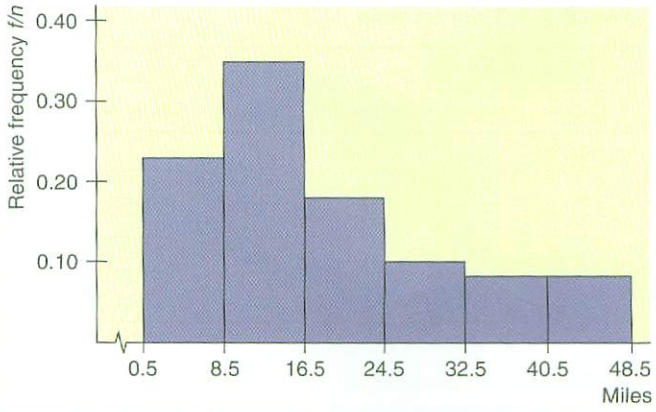


FIGURE 2-3

Relative-Frequency Histogram for Dallas Commuters: One-Way Commuting Distances



◇

◇ **COMMENT** The use of class boundaries in histograms assures us that the bars of the histogram touch and that no data fall on the boundaries. Both of these features are important. But a histogram displaying class boundaries may look awkward. For instance, the mileage range of 8.5 to 16.5 miles shown in Figure 2-2 isn't as natural a choice as a mileage range of 8 to 16 miles. For this reason, many magazines and newspapers do not use class boundaries as labels on a histogram. Instead, some use lower class limits as labels, with the convention that *a data value falling on the class limit is included in the next higher class (class to the right of the limit)*. Another convention is to label midpoints instead of class boundaries. Determine the convention being used before creating frequency tables and histograms on a computer. ◇

GUIDED EXERCISE 1

Histogram and relative-frequency histogram

One irate customer called Dollar Day Mail Order Company 40 times during the last two weeks to see why his order had not arrived. Each time he called, he recorded the length of time he was put "on hold" before being allowed to talk to a customer service representative.

TABLE 2-4 Length of Time on Hold, in Minutes

1	5	5	6	7	4	8	7	6	5
5	6	7	6	6	5	8	9	9	10
7	8	11	2	4	6	5	12	13	6
3	7	8	8	9	9	10	9	8	9

(a) What are the largest and smallest values in Table 2-4? If we want five classes in a frequency table, what should the class width be?

➡ The largest value is 13; the smallest value is 1. The class width is

$$\frac{13 - 1}{5} = 2.4 \approx 3$$ Note: *Increase* the value to 3.

(b) Complete the following frequency table.

TABLE 2-5 Time on Hold

Class Limits					
Lower–Upper		Tally	Frequency	Midpoint
1	– 3	_____	_____	_____
4	– _____	_____	_____	_____
_____	– 9	_____	_____	_____
_____	– _____	_____	_____	_____
_____	– _____	_____	_____	_____

➡

TABLE 2-6 Completion of Table 2-5

Class Limits			
Lower–Upper	Tally	Frequency	Midpoint
1–3	III	3	2
4–6	𝍸 𝍸 𝍸	15	5
7–9	𝍸 𝍸 𝍸 II	17	8
10–12	IIII	4	11
13–15	I	1	14

(c) Recall that the class boundary is halfway between the upper limit of one class and the lower limit of the next. Use this fact to find the class boundaries in Table 2-7 and to complete the partial histogram in Figure 2-4.

TABLE 2-7 Class Boundaries

Class Limits	Class Boundaries
1–3	0.5–3.5
4–6	3.5–6.5
7–9	6.5–_____
10–12	_____–_____
13–15	_____–_____

➡

TABLE 2-8 Completion of Table 2-7

Class Limits	Class Boundaries
1–3	0.5–3.5
4–6	3.5–6.5
7–9	6.5–9.5
10–12	9.5–12.5
13–15	12.5–15.5

Continued

GUIDED EXERCISE 1 continued

FIGURE 2-4

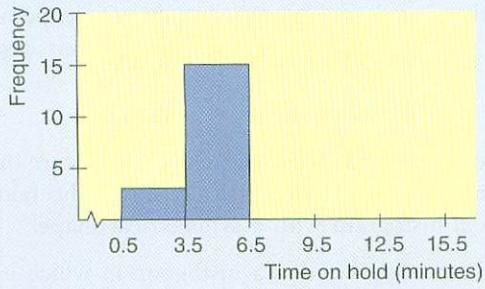
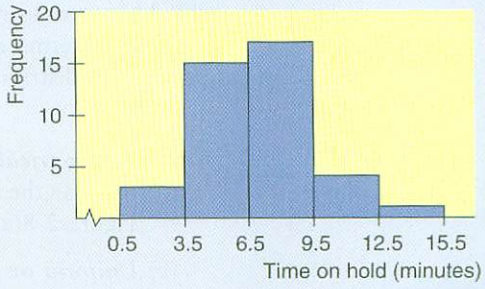


FIGURE 2-5 Completion of Figure 2-4



(d) Compute the relative class frequency f/n for each class in Table 2-9 and complete the partial relative-frequency histogram in Figure 2-6.

TABLE 2-9 Relative Class Frequency

Class	f/n
1–3	3/40 = 0.075
4–6	15/40 = 0.375
7–9	_____
10–12	_____
13–15	_____

TABLE 2-10 Completion of Table 2-9

Class	f/n
1–3	0.075
4–6	0.375
7–9	0.425
10–12	0.100
13–15	0.025

FIGURE 2-6

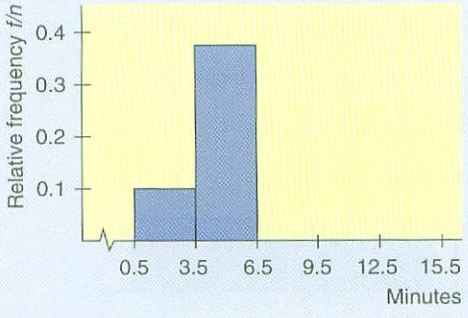
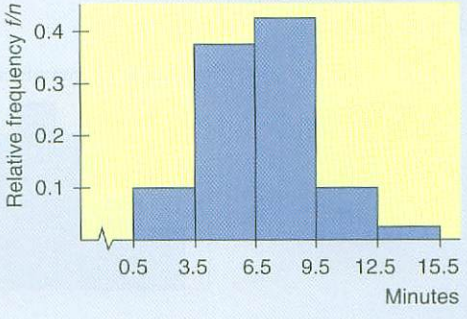


FIGURE 2-7 Completion of Figure 2-6



We will see relative-frequency distributions again when we study probability in Chapter 5. There we will see that if a random sample is large enough, then we can estimate the probability of an event by the relative frequency of the event. The relative-frequency distribution then can be interpreted as a *probability distribution.* Such distributions will form the basis of our work in inferential statistics.

Distribution Shapes

Histograms are valuable and useful tools. If the raw data came from a random sample of population values, the histogram constructed from the sample values should have a distribution shape that is reasonably similar to that of the population.

Several terms are commonly used to describe histograms and their associated population distributions.

(a) ***Symmetrical:*** This term refers to a histogram in which both sides are (more or less) the same when the graph is folded vertically down the middle. Figure 2-8(a) shows a typical histogram with a symmetrical shape.

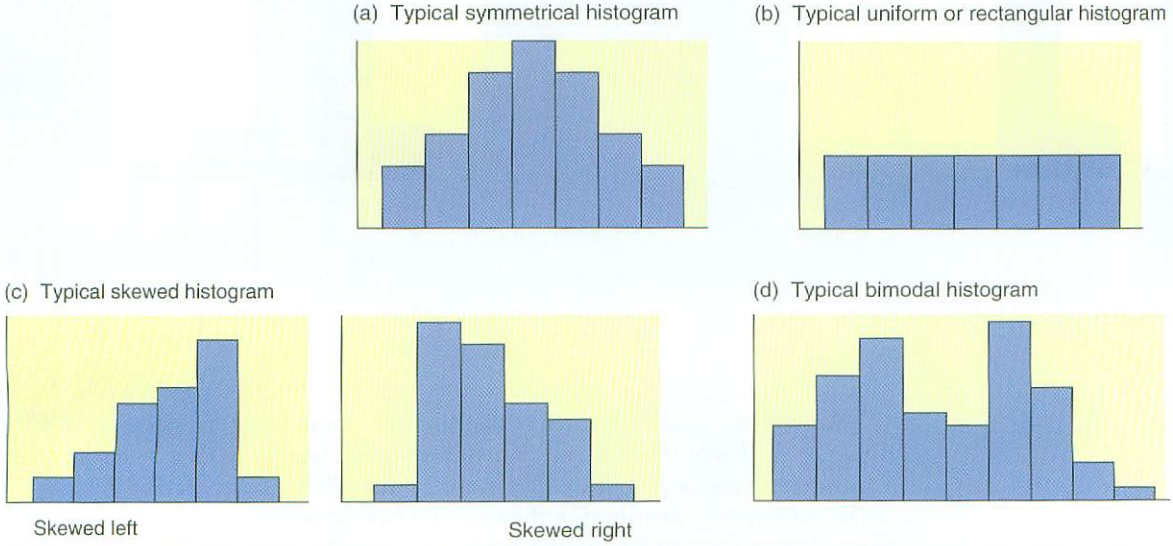
(b) ***Uniform or rectangular:*** These terms refer to a histogram in which every class has equal frequency. From one point of view, a uniform distribution is symmetrical with the added property that the bars are of the same height. Figure 2-8(b) illustrates a typical histogram with a uniform shape.

(c) ***Skewed left or skewed right:*** These terms refer to a histogram in which one tail is stretched out longer than the other. The direction of skewness is on the side of the *longer* tail. So if the longer tail is on the left, we say the histogram is skewed to the left. Figure 2-8(c) shows a typical histogram skewed to the left and another skewed to the right.

(d) ***Bimodal:*** This term refers to a histogram in which the two classes with the largest frequencies are separated by at least one class. The top two frequencies of these classes may have slightly different values. This type of situation sometimes indicates that we are sampling from two different populations. Figure 2-8(d) illustrates a typical histogram with a bimodal shape.

FIGURE 2-8

Types of Histograms



TECH NOTE The TI-84Plus/TI-83Plus calculators, Excel, and Minitab all create histograms. However, each technology automatically selects the number of classes to use. In Using Technology at the end of this chapter, you will see instructions for specifying the number of classes yourself and for generating histograms such as those we create "by hand."

VIEWPOINT

Mush, You Huskies!

In 1925, the village of Nome, Alaska, had a terrible diphtheria epidemic. Serum was available in Anchorage but had to be brought to Nome by dogsled over the 1161-mile Iditarod Trail. Since 1973, the Iditarod Dog Sled Race from Anchorage to Nome has been an annual sporting event with a current purse of more than $600,000. Winning times range from more than 20 days to a little over 9 days.

To collect data on winning times, visit the Brase/Brase statistics site at **http://math.college.hmco.com/students** and find the link to the Iditarod. Make a frequency distribution for these times.

SECTION 2.1 PROBLEMS

For Problems 1–6, use the specified number of classes to do the following:
(a) Find the class width.
(b) Make a frequency table showing class limits, class boundaries, midpoints, frequencies, and relative frequencies.
(c) Draw a histogram.
(d) Draw a relative-frequency histogram.

1. *Sports: Dog Sled Racing* How long does it take to finish the 1161-mile Iditarod Dog Sled Race from Anchorage to Nome, Alaska (*see* Viewpoint)? Finish times (to the nearest hour) for 57 dogsled teams are shown below.

261	271	236	244	279	296	284	299	288	288	247	256
338	360	341	333	261	266	287	296	313	311	307	307
299	303	277	283	304	305	288	290	288	289	297	299
332	330	309	328	307	328	285	291	295	298	306	315
310	318	318	320	333	321	323	324	327			

Use five classes.

2. *Medical: Glucose Testing* The following data represent glucose blood levels (mg/100 ml) after a 12-hour fast for a random sample of 70 women (Reference:

American Journal of Clinical Nutrition, Vol. 19, pp. 345–351). *Note:* These data are also available with other software on the statSpace CD-ROM.

45	66	83	71	76	64	59	59
76	82	80	81	85	77	82	90
87	72	79	69	83	71	87	69
81	76	96	83	67	94	101	94
89	94	73	99	93	85	83	80
78	80	85	83	84	74	81	70
65	89	70	80	84	77	65	46
80	70	75	45	101	71	109	73
73	80	72	81	63	74		

Use six classes.

3. *Medical: Tumor Recurrence* Certain kinds of tumors tend to recur. The following data represent the lengths of time, in months, for a tumor to recur after chemotherapy (Reference: D. P. Byar, *Journal of Urology*, Vol. 10, pp. 556–561). *Note:* These data are also available with other software on the statSpace CD-ROM.

19	18	17	1	21	22	54	46	25	49
50	1	59	39	43	39	5	9	38	18
14	45	54	59	46	50	29	12	19	36
38	40	43	41	10	50	41	25	19	39
27	20								

Use five classes.

4. *Archaeology: New Mexico* The Wind Mountain excavation site in New Mexico is an important archaeological location of the ancient Native American Anasazi culture. The following data represent depths (in cm) below surface grade at which significant artifacts were discovered at this site (Reference: Woosley, A. I. and McIntyre, A. J. *Mimbres Mogollon Archaeology*, University of New Mexico Press). *Note:* These data are also available with other software on the statSpace CD-ROM.

85	45	75	60	90	90	115	30	55	58
78	120	80	65	65	140	65	50	30	125
75	137	80	120	15	45	70	65	50	45
95	70	70	28	40	125	105	75	80	70
90	68	73	75	55	70	95	65	200	75
15	90	46	33	100	65	60	55	85	50
10	68	99	145	45	75	45	95	85	65
65	52	82							

Use seven classes.

5. *Environment: Gasoline Consumption* The following data represent highway fuel consumption in miles per gallon (mpg) for a random sample of 55 models of passenger cars (Source: Environmental Protection Agency). *Note:* These data are also available with other software on the statSpace CD-ROM.

30	27	22	25	24	25	24	15
35	35	33	52	49	10	27	18
20	23	24	25	30	24	24	24
18	20	25	27	24	32	29	27
24	27	26	25	24	28	33	30
13	13	21	28	37	35	32	33
29	31	28	28	25	29	31	

Use five classes.

6. *Advertising: Readability* "Readability Levels of Magazine Ads," by F. K. Shuptrine and D. D. McVicker, is an article in the *Journal of Advertising Research*. (For more information, visit the Brase/Brase statistics site at **http://math.college.hmco.com/students** and find the link to DASL, the Carnegie Mellon University Data and Story Library. Look in Data Subjects under Consumer and then Magazine Ads Readability file.) The following is a list of the number of three-syllable (or longer) words in advertising copy of randomly selected magazine advertisements.

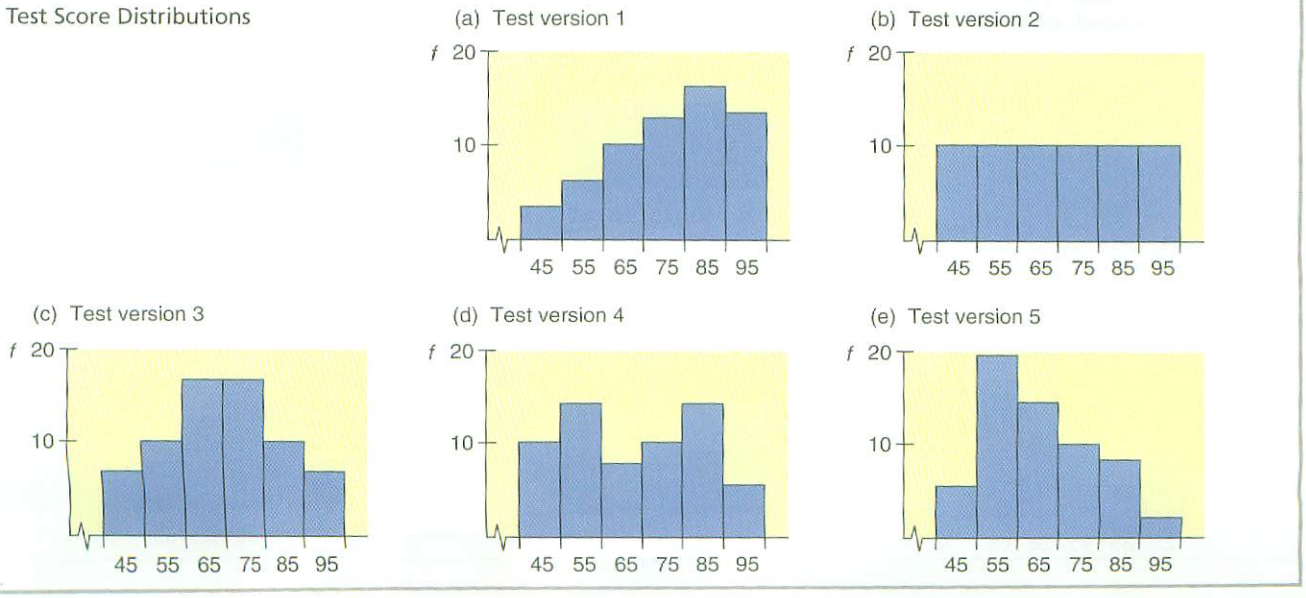
34	21	37	31	10	24	39	10	17	18	32
17	3	10	6	5	6	6	13	22	25	3
5	2	9	3	0	4	29	26	5	5	24
15	3	8	16	9	10	3	12	10	10	10
11	12	13	1	9	43	13	14	32	24	15

Use eight classes.

7. *Education: Testing* Professor Silva teaches anatomy and physiology. He has developed five different versions of a test on the same material. On giving each version to a different sample of 60 students, he discovered that the test score distributions looked like those shown in Figure 2-9.

FIGURE 2-9

Test Score Distributions



(a) Categorize the distribution shapes as uniform, symmetric, bimodal, skewed left, or skewed right.
(b) Comment on some advantages or problems with each test version. As a student, which version might you prefer? Which version would you like the least?

8. *Consumer: Warranty Cards* Many products come with owner registration or warranty cards. Usually, the consumer is asked a few questions about his or her family and household income. Random samples of warranty or registration cards for the indicated product revealed the household income distributions shown in Figure 2-10 below.
(a) Categorize the distribution shapes as uniform, symmetric, bimodal, skewed left, or skewed right.
(b) If you were in charge of advertising, how would you use income-distribution information of present customers to target ads for the indicated product?
(c) How valid do you think income information is on warranty cards?

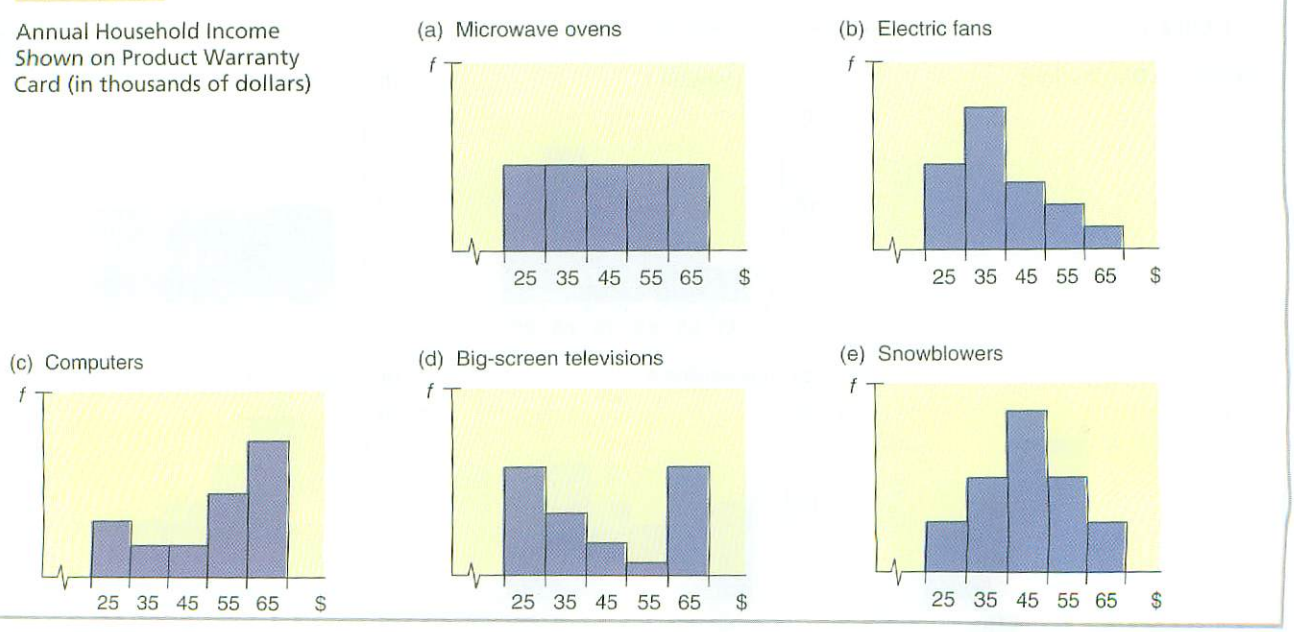
9. *Expand Your Knowledge: Decimal Data* The following data represent tonnes of wheat harvested each year (1894–1925) from Plot 19 at the Rothamsted Agricultural Experiment Stations, England.

2.71	1.62	2.60	1.64	2.20	2.02	1.67	1.99	2.34	1.26	1.31
1.80	2.82	2.15	2.07	1.62	1.47	2.19	0.59	1.48	0.77	2.04
1.32	0.89	1.35	0.95	0.94	1.39	1.19	1.18	0.46	0.70	

(a) Multiply each data value by 100 to "clear" the decimals.
(b) Use the standard procedures of this section to make a frequency table and histogram with your whole-number data. Use six classes.
(c) Divide class limits, class boundaries, and class midpoints by 100 to get back to your original data values.

FIGURE 2-10

Annual Household Income Shown on Product Warranty Card (in thousands of dollars)



10. *Decimal Data: Batting Averages* The following data represent baseball batting averages for a random sample of National League players near the end of the baseball season. The data are from the baseball statistics section of *The Denver Post*.

0.194	0.258	0.190	0.291	0.158	0.295	0.261	0.250	0.181
0.125	0.107	0.260	0.309	0.309	0.276	0.287	0.317	0.252
0.215	0.250	0.246	0.260	0.265	0.182	0.113	0.200	

(a) Multiply each data value by 1000 to "clear" the decimals.
(b) Use the standard procedures of this section to make a frequency table and histogram with your whole-number data. Use five classes.
(c) Divide class limits, class boundaries, and class midpoints by 1000 to get back to your original data.

11. *Expand Your Knowledge: Dotplot* Another display technique that is somewhat similar to a histogram is a *dotplot*. In a dotplot, the data values are displayed along the horizontal axis. A dot is then plotted over each data value in the data set.

PROCEDURE

How to make a dotplot

Display the data along a horizontal axis. Then plot each data value with a dot or point above the corresponding value on the horizontal axis. For repeated data values, stack the dots.

The next display shows a dotplot generated by Minitab (➤**Graph** ➤**Dotplot**) for the number of licensed drivers per 1000 residents by state, including the District of Columbia (Source: U.S. Department of Transportation).

Dotplot for Licensed Drivers per 1000 Residents

(a) From the dotplot, how many states have 600 or fewer licensed drivers per 1000 residents?
(b) About what percentage of the states (out of 51) seem to have close to 800 licensed drivers per 1000 residents?
(c) Consider the intervals 550 to 650, 650 to 750, and 750 to 850 licensed drivers per 1000 residents. In which interval do most of the states fall?

12. *Dotplot: Dog Sled Racing* Make a dotplot for the data in Problem 1 regarding the finish time (number of hours) for the Iditarod Dog Sled Race. Compare the dotplot to the histogram of Problem 1.

13. *Dotplot: Tumor Recurrence* Make a dotplot for the data in Problem 3 regarding the recurrence of tumors after chemotherapy. Compare the dotplot to the histogram of Problem 3.

2.2 Bar Graphs, Circle Graphs, and Time-Series Graphs

FOCUS POINTS

- ✓ Determine types of graphs appropriate for specific data.
- ✓ Construct bar graphs, Pareto charts, circle graphs, and time-series graphs.
- ✓ Interpret information displayed in graphs.

Histograms provide a useful visual display of the distribution of data. However, the data must be quantitative. In this section, we examine other types of graphs, some of which are suitable for qualitative or category data as well.

Let's start with *bar graphs*. These are graphs that can be used to display quantitative or qualitative data.

Bar graphs

Features of a bar graph

1. Bars can be vertical or horizontal.
2. Bars are of uniform width and uniformly spaced.
3. The lengths of the bars represent values of the variable being displayed, the frequency of occurrence, or the percentage of occurrence. The same measurement scale is used for the length of each bar.
4. The graph is well annotated with title, labels for each bar, and vertical scale or actual value for the length of each bar.

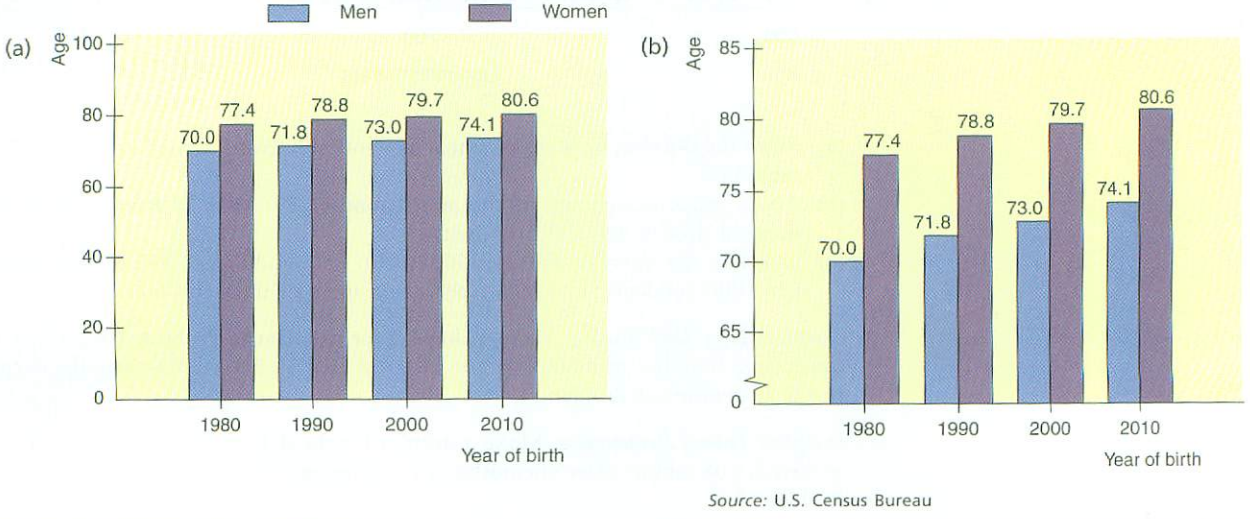
EXAMPLE 3 *Bar graph*

Figure 2-11 shows two bar graphs depicting the life expectancies for men and women born in the designated year. Let's analyze the features of these graphs.

SOLUTION: The graphs are called *clustered bar graphs* because there are two bars for each year of birth. One bar represents the life expectancy for men, and the other represents the life expectancy for women. The height of each bar represents the life expectancy (in years). ◆

FIGURE 2-11

Life Expectancy



Source: U.S. Census Bureau

Changing scales

An important feature illustrated in Figure 2-11(b) is that of a *changing scale*. Notice that the scale between 0 and 65 is compressed. The changing scale amplifies the apparent difference between life spans for men and women, as well as the increase in life spans from those born in 1980 to the projected span of those born in 2010.

Changing Scale

Whenever you use a change in scale in a graphic, warn the viewer by using a squiggle ⌇ on the changed axis. Sometimes, if a single bar is unusually long, the bar length is compressed with a squiggle in the bar itself.

Pareto charts

Quality control is an important aspect of today's production and service industries. Dr. W. Edwards Deming was one of the developers of total quality management (TQM). In his book *Out of Crisis*, he outlines many strategies for monitoring and improving service and production industries. In particular, Dr. Deming recommends the use of some statistical methods to organize and analyze data from industries so that sources of problems can be identified and then corrected. *Pareto* (pronounced "Pah-rāy-tō) *charts* are among the many techniques used in quality-control programs.

A **Pareto chart** is a bar graph in which the bar height represents frequency of an event. In addition, the bars are arranged from left to right according to decreasing height.

GUIDED EXERCISE 2

Pareto charts

This exercise is adapted from *The Deming Management Method* by Mary Walton. Suppose you want to arrive at college 15 minutes before your first class so that you can feel relaxed when you walk into class. An early arrival time also allows room for unexpected delays. However, you always find yourself arriving "just in time" or slightly late. What causes you to be late? Charlotte made a list of possible causes and then kept a checklist for 2 months (Table 2-11). On some days more than one item was checked because several events occurred that caused her to be late.

Continued

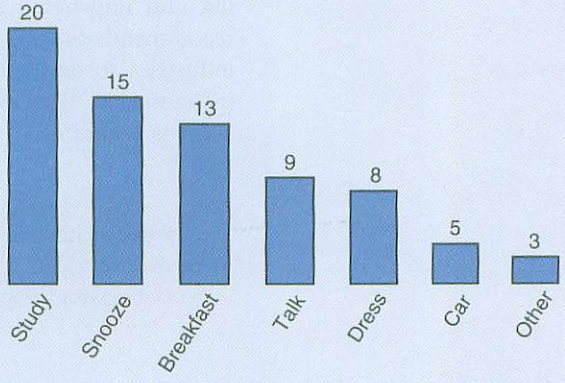
GUIDED EXERCISE 2 continued

TABLE 2-11 Causes for Lateness
(September–October)

Cause	Frequency
Snoozing after alarm goes off	15
Car trouble	5
Too long over breakfast	13
Last-minute studying	20
Finding something to wear	8
Talking too long with roommate	9
Other	3

(a) Make a Pareto chart showing the causes for lateness. Be sure to label the causes, and draw the bars using the same vertical scale.

FIGURE 2-12 Pareto Chart: Conditions That Might Cause Lateness



(b) Looking at the Pareto chart, what recommendations do you have for Charlotte?

According to the chart, rearranging study time, or getting up earlier to allow for studying, would cure her most frequent cause for lateness. Repairing the car might be important, but for getting to campus early, it would not be as effective as adjusting study time.

Circle graphs or pie charts

Another popular pictorial representation of data is the *circle graph* or *pie chart*. It is relatively safe from misinterpretation and is especially useful for showing the division of a total quantity into its component parts. The total quantity, or 100%, is represented by the entire circle. Each wedge of the circle represents a component part of the total. These proportional segments are usually labeled with corresponding percentages of the total. Guided Exercise 3 shows how to make a circle graph.

In a **circle graph** or **pie chart,** wedges of a circle visually display proportional parts of the total population that share a common characteristic.

GUIDED EXERCISE 3

Circle graph

How long do we spend talking on the telephone after hours (at home after 5 P.M.)? The results from a recent survey of 500 people (as reported in *USA Today*) are shown in Table 2-12. We'll make a circle graph to display these data.

TABLE 2-12 Time Spent on Home Telephone After 5 P.M.

Time	Number	Fractional Part	Percentage	Number of Degrees
Less than ½ hour	296	296/500	59.2	59.2% × 360° ≈ 213°
½ hour to 1 hour	83	83/500	16.6	16.6% × 360° ≈ 60°
More than 1 hour	121	____	____	____
Total	____		____	____

(a) Fill in the missing parts in Table 2-12 for "More than 1 hour." Remember that the central angle of a circle is 360°. Round to the nearest degree.

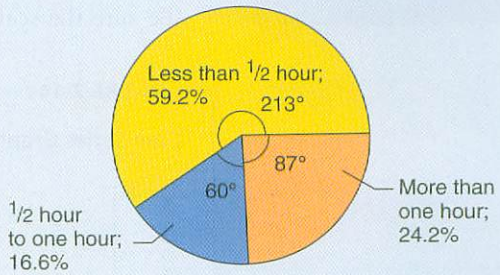
➡ For "More than 1 hour," Fractional Part = 121/500; Percentage = 24.2%; Number of Degrees = 24.2% × 360° ≈ 87°. The symbol ≈ means approximately equal.

(b) Fill in the totals. What is the total number of responses? Do the percentages total 100% (within rounding error)? Do the numbers of degrees total 360° (within rounding error)?

➡ The total number of responses is 500. The percentages total 100%. You must have such a total in order to create a circle graph. The numbers of degrees total 360°.

(c) Draw a circle graph. Divide the circle into pieces with the designated numbers of degrees. Label each piece, and show the percentage corresponding to each piece. The numbers of degrees are usually omitted from pie charts shown in newspapers, magazines, journals, and reports.

➡ **FIGURE 2-13** Hours on Home Telephone After 5 P.M.



Suppose you begin an exercise program that involves walking or jogging for 30 minutes. You exercise several times a week but monitor yourself by logging the distance you cover in 30 minutes each Saturday. How do you display these data in a meaningful way? Making a bar chart showing the frequency of distances you cover might be interesting, but it does not really show how the distance you cover in 30 minutes has changed over time. A graph showing the distance covered on each date will let you track your performance over time.

Time-series graph

We will use a *time-series graph*. A time-series graph is a graph showing data measurements in chronological order. To make a time-series graph, we put time on the horizontal scale and the variable being measured on the vertical scale. In a basic time-series graph, we connect the data points by lines.

In a **time-series graph,** data are plotted in order of occurrence at regular intervals over a period of time.

EXAMPLE 4
Time-series graph

Suppose you have been in the walking/jogging exercise program for 20 weeks, and for each week you have recorded the distance you covered in 30 minutes. Your data log is shown in Table 2-13.

TABLE 2-13 Distance (in Miles) Walked/Jogged in 30 Minutes

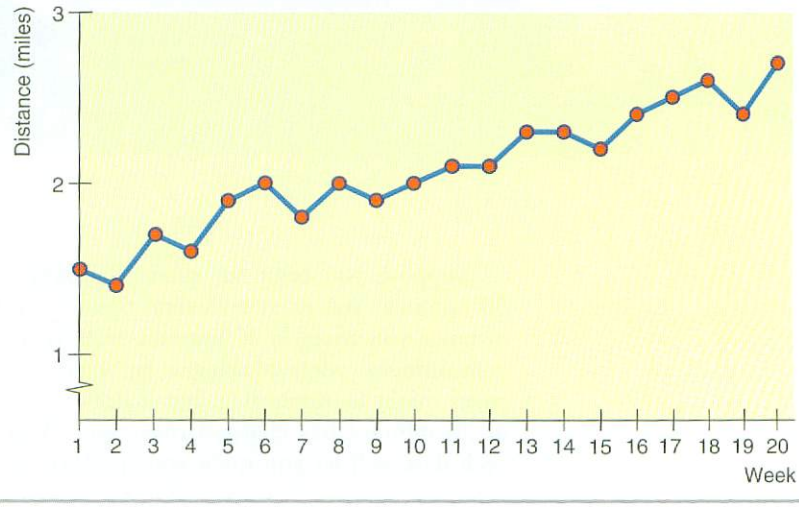
Week	1	2	3	4	5	6	7	8	9	10
Distance	1.5	1.4	1.7	1.6	1.9	2.0	1.8	2.0	1.9	2.0
Week	11	12	13	14	15	16	17	18	19	20
Distance	2.1	2.1	2.3	2.3	2.2	2.4	2.5	2.6	2.4	2.7

(a) Make a time-series graph.

SOLUTION: The data are appropriate for a time-series graph because they represent the same measurement (distance covered in a 30-minute period) taken at different times. The measurements are also recorded at equal time intervals (every week). To make our time-series graph, we list the weeks in order on the horizontal scale. Above each week, plot the distance covered that week on the vertical scale. Then connect the dots. Figure 2-14 shows the time-series graph. Be sure the scales are labeled.

FIGURE 2-14

Time-Series Graph of Distance (in miles) Jogged in 30 Minutes



(b) From looking at Figure 2-14, can you detect any patterns?

SOLUTION: There seems to be an upward trend in distance covered. The distances covered in the last few weeks are about a mile farther than those for the first few weeks. However, we cannot conclude that this trend will continue. Perhaps you have reached your goal for this training activity and now wish to maintain a distance of about 2.5 miles in 30 minutes. ◇

Time series

Data sets composed of similar measurements taken at regular intervals over time are called *time series*. Time series are often used in economics, finance, sociology, medicine, and any situation in which we want to study or monitor a similar measure over a period of time. A time-series graph can reveal some of the main features of time series.

Time-series data consist of measurements of the same variable for the same subject taken at regular intervals over a period of time.

We've seen several styles of graphs. Which kinds are suitable for a specific data collection?

PROCEDURE

How to decide which type of graph to use

Bar graphs are useful for quantitative or qualitative data. With qualitative data, the frequency or percentage of occurrence can be displayed. With quantitative data, the measurement itself can be displayed, as was done in the bar graph showing life expectancy. Watch that the measurement scale is consistent or that a jump scale squiggle is used.

Pareto charts identify the frequency of events or categories in decreasing order of frequency of occurrence.

Circle graphs display how a *total* is dispersed into several categories. The circle graph is very appropriate for qualitative data, or any data for which percentage of occurrence makes sense. Circle graphs are most effective when the number of wedges is 10 or fewer.

Time-series graphs display how data change over time. It is best if the units of time are consistent in a given graph. For instance, measurements taken every day should not be mixed on the same graph with data taken every week.

For any graph: Provide a title, label the axes, and identify units of measure. As Edward Tufte suggests in his book *The Visual Display of Quantitative Information*, don't let artwork or skewed perspective cloud the clarity of the information displayed.

TECH NOTES *Bar graphs, circle graphs, and time-series graphs*

TI-84Plus/TI-83Plus Only time-series. Place consecutive values 1 through the number of time segments in list L1 and corresponding data in L2. Press **Stat Plot** and highlight an *xy* line plot.

Excel Use the chart wizard on the toolbar . Select the desired option and follow the instructions in the dialogue boxes.

Minitab Use the menu selection **Graph.** Select the desired option and follow the instructions in the dialogue boxes.

VIEWPOINT

Do Ethical Standards Vary by the Situation?

The Lutheran Brotherhood did a national survey and found that nearly 60% of all U.S. adults claim that ethics vary by the situation; 33% claim that there is only one ethical standard; and 7% were not sure. How could you draw a circle graph to make a visual impression of Americans' views on ethical standards?

SECTION 2.2 PROBLEMS

1. *Education: Does College Pay Off?* It is costly in both time and money to go to college. Does it pay off? According to the Bureau of Census, the answer is yes. The average annual income (in thousands of dollars) of a *household* headed by a person with the stated education level is as follows: 16.1 if ninth grade is the highest level achieved, 34.3 for high school graduates, 48.6 for those holding associate degrees, 62.1 for those with bachelor's degrees, 71.0 for those with master's degrees, and 84.1 for those with doctoral degrees. Make a bar graph showing household income for each education level.

2. *Accidents: Child Deaths* How safe is the world for kids? Unfortunately, some children between the ages of 1 and 14 die of injuries every year. United Nations data show that by nation, the annual numbers of deaths from injuries per 100,000 children are as follows: Australia, 9.5; Canada, 9.7; Denmark, 8.1; France, 9.1; Germany, 8.3; Hungary, 10.8; Ireland, 8.3; Italy, 6.1; Japan, 8.4; Korea, 25.6; Mexico, 19.8; New Zealand, 13.7; Netherlands, 6.6; Poland, 13.4; Portugal, 17.8; Spain, 8.1; Sweden, 5.2; Switzerland, 9.6; U.K., 6.1; United States, 14.1. Display these data in a Pareto chart.

3. *Commercial Fishing: Gulf of Alaska* It's not an easy life, but it's a good life! Suppose you decide to take the summer off and sign on as a deck hand for a commercial fishing boat in Alaska that specializes in deep-water fishing for groundfish.

What kind of fish can you expect to catch? One way to answer this question is to examine government reports on groundfish caught in the Gulf of Alaska. The following list indicates the types of fish caught annually in thousands of metric tons (Source: *Report on the Status of U.S. Living Marine Resources*, National Oceanic and Atmospheric Administration): flatfish, 36.3; Pacific cod, 68.6; sablefish, 16.0; Walleye pollock, 71.2; rockfish, 18.9. Make a Pareto chart showing the annual harvest for commercial fishing in the Gulf of Alaska.

4. *Archaeology: Ireland* Commercial dredging operations in ancient rivers occasionally uncover archaeological artifacts of great importance. One such artifact is Bronze Age spearheads recovered from ancient rivers in Ireland. A recent study gave the following information regarding discoveries of ancient bronze spearheads in Irish rivers.

River	Bann	Blackwater	Erne	Shannon	Barrow
No. of spearheads	19	8	15	33	14

(Based on information from *Crossing the Rubicon, Bronze Age Studies 5*, Lorraine Bourke, Department of Archaeology, National University of Ireland, Galway.)
(a) Make a Pareto chart for these data.
(b) Make a circle graph for these data.

5. *Lifestyle: Hide the Mess!* A survey of 1000 adults (reported in *USA Today*) uncovered some interesting housekeeping secrets. When unexpected company comes, where do we hide the mess? The survey showed that 68% of the respondents toss their mess in the closet, 23% shove things under the bed, 6% put things in the bathtub, and 3% put the mess in the freezer. Make a circle graph to display this information.

6. *Education: College Professors' Time* How do college professors spend their time? *The National Education Association Almanac of Higher Education* gives the following average distribution of professional time allocation: teaching, 51%; research, 16%; professional growth, 5%; community service, 11%; service to the college, 11%; and consulting outside the college, 6%. Make a pie chart showing the allocation of professional time for college professors.

7. *FBI Report: Hawaii* In the Aloha state, you are very unlikely to be murdered! However, it is considerably more likely that your house might be burgled, your car might be stolen, or you might be punched in the nose. That said, Hawaii is still a great place for a vacation or, if you are very lucky, to live. The following numbers represent the crime rates per 100,000 population in Hawaii: murder, 2.6; rape, 33.4; robbery, 93.3; house burglary, 911.6; motor vehicle theft, 550.7; assault, 125.3 (Source: *Crime in the United States*, U.S. Department of Justice, Federal Bureau of Investigation).
   (a) Display this information in a Pareto chart, showing the crime rate for each category.
   (b) Could the information as reported be displayed as a circle graph? Explain. *Hint:* Other forms of crime, such as arson, are not included in the information. In addition, some crimes might occur together.

8. *Driving: Bad Habits* Driving would be more pleasant if we didn't have to put up with the bad habits of other drivers. *USA Today* reported the results of a Valvoline Oil Company survey of 500 drivers in which the drivers marked their complaints about other drivers. The top complaints turned out to be tailgating, marked by 22% of the respondents; not using turn signals, marked by 19%; 16% marked being cut off; 11% complained about other drivers driving too slowly; and 8% complained

about other drivers being inconsiderate. Make a Pareto chart showing percentage of drivers listing each stated complaint. Could this information as reported be put in a circle graph? Why or why not?

9. *Ecology: Lakes* Pyramid Lake, Nevada, is described as the pride of the Paiute Indian Nation. It is a beautiful desert lake famous for very large trout. The elevation of the lake surface (feet above sea level) varies according to the annual flow of the Truckee River from Lake Tahoe. The U.S. Geological Survey provided the following data:

Year	Elevation
1986	3817
1987	3815
1988	3810
1989	3812
1990	3808
1991	3803

Year	Elevation
1992	3798
1993	3797
1994	3795
1995	3797
1996	3802
1997	3807

Year	Elevation
1998	3811
1999	3816
2000	3817

Make a time-series graph displaying the data. For more information, visit the Brase/Brase statistics site at **http://math.college.hmco.com/students** and find the link to the Pyramid Lake Fisheries.

10. *Vital Statistics: Height* How does average height for boys change as boys get older? According to *Physician's Handbook*, the average heights at different ages are as follows:

Age (years)	0.5	1	2	3	4	5	6	7
Height (inches)	26	29	33	36	39	42	45	47

Age (years)	8	9	10	11	12	13	14
Height (inches)	50	52	54	56	58	60	62

Make a time-series graph for average height for ages 0.5 through 14 years.

2.3 Stem-and-Leaf Displays

Exploratory Data Analysis

FOCUS POINTS

- ✓ Construct a stem-and-leaf display from raw data.
- ✓ Use a stem-and-leaf display to visualize data distribution.
- ✓ Compare a stem-and-leaf display to a histogram.

Together with histograms and other graphics techniques, the stem-and-leaf display is one of many useful ways of studying data in a field called *exploratory data analysis* (often abbreviated as *EDA*). John W. Tukey wrote one of the definitive books on the subject, *Exploratory Data Analysis* (Addison-Wesley). Another very useful reference for EDA techniques is the book *Applications, Basics, and Computing of Exploratory Data Analysis* by Paul F. Velleman and David C. Hoaglin

(Duxbury Press). Exploratory data analysis techniques are particularly useful for detecting patterns and extreme data values. They are designed to help us explore a data set, to ask questions we had not thought of before, or to pursue leads in many directions.

EDA techniques are similar to those of an explorer. An explorer has a general idea of destination but is always alert to the unexpected. An explorer needs to assess situations quickly and often simplify and clarify them. An explorer makes pictures—that is, maps showing the relationships of landscape features. The aspects of rapid implementation, visual displays such as graphs and charts, data simplification, and robustness (that is, analysis that is not influenced much by extreme data values) are key ingredients of EDA techniques. In addition, these techniques are good for exploration because they require very few prior assumptions about the data.

EDA methods are especially useful when our data have been gathered for general interest and observation of subjects. For example, we may have data regarding the ages of applicants to graduate programs. We don't have a specific question in mind. We want to see what the data reveal. Are the ages fairly uniform or spread out? Are there exceptionally young or old applicants? If there are, we might look at other characteristics of these applicants, such as field of study. EDA methods help us quickly absorb some aspects of the data and then may lead us to ask specific questions to which we might apply methods of traditional statistics.

In contrast, when we design an experiment to produce data to answer a specific question, we focus on particular aspects of the data that are useful to us. If we want to determine the average highway gas mileage of a specific sports car, we use that model car in well-designed tests. We don't need to worry about unexpected road conditions, poorly trained drivers, different fuel grades, sudden stops and starts, etc. Our experiment is designed to control outside factors. Consequently, we do not need to "explore" our data as much. We can often make valid assumptions about the data. Methods of traditional statistics will be very useful to analyze such data and answer our specific questions.

Stem-and-Leaf Display

In this text, we will introduce two EDA techniques: stem-and-leaf displays and, in Section 3.3, box-and-whisker plots. Let's first look at a stem-and-leaf display.

A **stem-and-leaf display** is a method of exploratory data analysis that is used to rank-order and arrange data into groups.

We know that frequency distributions and histograms provide a useful organization and summary of data. However, in a histogram, we lose most of the specific data values. A stem-and-leaf display is a device that organizes and groups data but allows us to recover the original data if desired. In the next example, we will make a stem-and-leaf display.

EXAMPLE 5
Stem-and-leaf display

Many airline passengers seem weighted down with their carry-on luggage. Just how much weight are they carrying? The carry-on luggage weights in pounds for a random sample of 40 passengers returning from a vacation to Hawaii were recorded (see Table 2-14).

TABLE 2-14 Weights of Carry-On Luggage in Pounds

30	27	12	42	35	47	38	36	27	35
22	17	29	3	21	0	38	32	41	33
26	45	18	43	18	32	31	32	19	21
33	31	28	29	51	12	32	18	21	26

To make a stem-and-leaf display, we break the digits of each data value into *two* parts. The left group of digits is called a *stem*, and the remaining group of digits on the right is called a *leaf*. We are free to choose the number of digits to be included in the stem.

The weights in our example consist of two-digit numbers. For a two-digit number, the stem selection is obviously the left digit. In our case, the tens digits will form the stems, and the units digits will form the leaves. For example, for the weight 12, the stem is 1, and the leaf is 2. For the weight 18, the stem is again 1, but the leaf is 8. In the stem-and-leaf display, we list each possible stem once on the left and all its leaves in the same row on the right, as in Figure 2-15(a). Finally, we order the leaves as shown in Figure 2-15(b).

Figure 2-15 shows a stem-and-leaf display for the weights of carry-on luggage. From the stem-and-leaf display in Figure 2-15, we see that two bags weighed 27 lb, one weighed 3 lb, one weighed 51 lb, and so on. We see that most of the weights were in the 30-lb range, only two were less than 10 lb, and six were over 40 lb. Note that the lengths of the lines containing the leaves give the visual impression that a sideways histogram would present.

As a final step, we need to indicate the scale. This is usually done by indicating the value represented by the stem and one leaf. ◆

There are no firm rules for selecting the group of digits for the stem. But whichever group you select, you must list all the possible stems from smallest to largest in the data collection.

FIGURE 2-15

Stem-and-Leaf Displays of Airline Carry-On Luggage Weights

(a) Leaves Not Ordered

3 | 2 represents 32 lb

Stem	Leaves
0	3 0
1	2 7 8 8 9 2 8
2	7 7 2 9 1 6 1 8 9 1 6
3	0 5 8 6 5 8 2 3 2 1 2 3 1 2
4	2 7 1 5 3
5	1

(b) Final Display with Leaves Ordered

3 | 2 represents 32 lb

Stem	Leaves
0	0 3
1	2 2 7 8 8 8 9
2	1 1 1 2 6 6 7 7 8 9 9
3	0 1 1 2 2 2 2 3 3 5 5 6 8 8
4	1 2 3 5 7
5	1

PROCEDURE

How to make a stem-and-leaf display

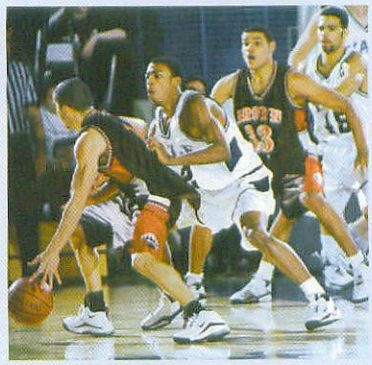
1. Divide the digits of each data value into two parts. The leftmost part is called the *stem* and the rightmost part is called the *leaf*.
2. Align all the stems in a vertical column from smallest to largest. Draw a vertical line to the right of all the stems.
3. Place all the leaves with the same stem in the same row as the stem, and arrange the leaves in increasing order.
4. Use a label to indicate the magnitude of the numbers in the display. We include the decimal position in the label rather than with the stems or leaves.

GUIDED EXERCISE 4

Stem-and-leaf display

What does it take to win at sports? If you're talking about basketball, one sports writer gave the answer. He listed the winning scores of the conference championship games over the last 35 years. The scores for those games follow below.

132	118	124	109	104	101	125	83	99
131	98	125	97	106	112	92	120	103
111	117	135	143	112	112	116	106	117
119	110	105	128	112	126	105	102	



To make a stem-and-leaf display, we'll use the first *two* digits as the stems (see Figure 2-16). Notice that the distribution of scores is fairly symmetrical.

(a) Use the first *two* digits as the stem. Then order the leaves. Provide a label that shows the meaning and units of the first stem and first leaf.

➡

FIGURE 2-16 Winning Scores

08	3	3 represents 083 or 83 points
08	3	
09	2 7 8 9	
10	1 2 3 4 5 5 6 6 9	
11	0 1 2 2 2 2 6 7 7 8 9	
12	0 4 5 5 6 8	
13	1 2 5	
14	3	

(b) Looking at the distribution, would you say that it is fairly symmetrical?

➡ Yes. Notice that stem 11 has the most data.

◇ **COMMENT** Stem-and-leaf displays organize the data, let the data analyst spot extreme values, and are easy to create. In fact, they can be used to organize data so that frequency tables are easier to make. However, at this time, histograms are used more often in formal data presentations, whereas stem-and-leaf displays are used by data analysts to gain initial insights about the data. ◇

TECH NOTE *Stem-and-Leaf Display*

TI-84Plus/TI-83Plus Does not support stem-and-leaf displays. You can sort the data by using keys **Stat ➤ Edit ➤ 2:SortA.**

Excel Does not support stem-and-leaf displays. You can sort the data by using menu choices **Data ➤ Sort.**

Minitab Use the menu selections **Graph ➤ Stem-and-Leaf** and fill in the dialogue box.

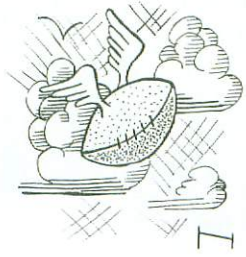
Minitab Release 14 Stem-and-Leaf Display (for Data in Guided Exercise 4)

```
Stem-and-Leaf of Scores     N=35
Leaf Unit=1.0

  1                8     3
  5                9     2789
 14               10     123455669
(11)              11     01222267789
 10               12     045568
  4               13     125
  1               14     3
```

The values shown in the left column represent depth. Numbers above the value in parentheses show the cumulative number of values from the top to the stem of the middle value. Numbers below the value in parentheses show the cumulative number of values from the bottom to the stem of the middle value. The number in parentheses shows how many values are on the same line as the middle value.

VIEWPOINT



What Does It Take to Win?

Scores for NFL Super Bowl games can be found at the NFL web site. Visit the Brase/Brase statistics site at **http://math.college.hmco.com/students** and find the link to the NFL. Once at the NFL web site, follow the links to the Super Bowl. Of special interest in football statistics is the spread, or difference, between scores of the winning and losing teams. If the spread is too large, the game can appear to be lopsided, and TV viewers become less interested in the game (and accompanying commercial ads). Make a stem-and-leaf display of the spread for the NFL Super Bowl games and analyze the results.

SECTION 2.3 PROBLEMS

1. *Cowboys: Longevity* How long did *real* cowboys live? One answer may be found in the book *The Last Cowboys* by Connie Brooks (University of New Mexico Press). This delightful book presents a thoughtful sociological study of cowboys in West Texas and Southeastern New Mexico around the year 1890. A sample of 32 cowboys gave the following years of longevity:

58	52	68	86	72	66	97	89	84	91	91
92	66	68	87	86	73	61	70	75	72	73
85	84	90	57	77	76	84	93	58	47	

   (a) Make a stem-and-leaf display for these data.
   (b) Consider the following quote from Baron von Richthofen in his *Cattle Raising on the Plains of North America:* "Cowboys are to be found among the sons of the best families. The truth is probably that most were not a drunken, gambling lot, quick to draw and fire their pistols." Does the data distribution of longevity lend credence to this quote?

2. *Ecology: Habitat* Wetlands offer a diversity of benefits. They provide habitat for wildlife, spawning grounds for U.S. commercial fish, and renewable timber resources. In the last 200 years the United States has lost more than half its wetlands. *Environmental Almanac* gives the percentage of wetlands lost in each state in the last 200 years. For the lower 48 states, the percentage loss of wetlands per state is as follows:

46	37	36	42	81	20	73	59	35	50
87	52	24	27	38	56	39	74	56	31
27	91	46	9	54	52	30	33	28	35
35	23	90	72	85	42	59	50	49	
48	38	60	46	87	50	89	49	67	

   Make a stem-and-leaf display of these data. Be sure to indicate the scale. How are the percentages distributed? Is the distribution skewed? Are there any gaps?

3. *Health Care: Hospitals* The American Medical Association Center for Health Policy Research included data, by state, on the number of community hospitals and the average patient stay (in days) in its publication *State Health Care Data: Utilization, Spending, and Characteristics.* The data (by state) are shown in the table. Make a stem-and-leaf display of the data for the average length of stay in days. Comment about the general shape of the distribution.

State	No. of Hospitals	Average Length of Stay	State	No. of Hospitals	Average Length of Stay	State	No. of Hospitals	Average Length of Stay
Alabama	119	7.0	Colorado	71	6.8	Georgia	162	7.2
Alaska	16	5.7	Connecticut	35	7.4	Hawaii	19	9.4
Arizona	61	5.5	Delaware	8	6.8	Idaho	41	7.1
Arkansas	88	7.0	Dist. of Columbia	11	7.5	Illinois	209	7.3
California	440	6.0	Florida	227	7.0	Indiana	113	6.6

Continued

State	No. of Hospitals	Average Length of Stay	State	No. of Hospitals	Average Length of Stay	State	No. of Hospitals	Average Length of Stay
Iowa	123	8.4	Nebraska	90	9.6	Rhode Island	12	6.9
Kansas	133	7.8	Nevada	21	6.4	S. Carolina	68	7.1
Kentucky	107	6.9	New Hampshire	27	7.0	S. Dakota	52	10.3
Louisiana	136	6.7	New Jersey	96	7.6	Tennessee	122	6.8
Maine	38	7.2	New Mexico	37	5.5	Texas	421	6.2
Maryland	51	6.8	New York	231	9.9	Utah	42	5.2
Massachusetts	101	7.0	N. Carolina	117	7.3	Vermont	15	7.6
Michigan	175	7.3	N. Dakota	47	11.1	Virginia	98	7.0
Minnesota	148	8.7	Ohio	193	6.6	Washington	92	5.6
Mississippi	102	7.2	Oklahoma	113	6.7	W. Virginia	59	7.1
Missouri	133	7.4	Oregon	66	5.3	Wisconsin	129	7.3
Montana	53	10.0	Pennsylvania	236	7.5	Wyoming	27	8.5

4. *Health Care: Hospitals* Using the number of hospitals per state listed in the table in Problem 3, make a stem-and-leaf display for the number of community hospitals per state. Which states have an unusually high number of hospitals?

5. *Expand Your Knowledge: Split Stem* The Boston Marathon is the oldest and best known U.S. marathon. It covers a route from Hopkinton, Massachusetts, to downtown Boston. The distance is approximately 26 miles. Visit the Brase/Brase statistics site at **http://math.college.hmco.com/students** and find the link to the Boston Marathon. Search the marathon site to find a wealth of information about the history of the race. In particular, the site gives the winning times for the Boston Marathon. They are all over 2 hours. The following data are the minutes over 2 hours for the winning male runners:

1961–1980

23	23	18	19	16	17	15	22	13	10
18	15	16	13	9	20	14	10	9	12

1981–2000

9	8	9	10	14	7	11	8	9	8
11	8	9	7	9	9	10	7	9	9

(a) Make a stem-and-leaf display for the minutes over 2 hours of the winning times for the years 1961 to 1980. Use two lines per stem.

PROCEDURE

How to split a stem

When a stem has many leaves, it is useful to split the stem into two lines or more. For two lines per stem, place leaves 0 to 4 on the first line and leaves 5 to 9 on the next line.

(b) Make a stem-and-leaf display for the minutes over 2 hours of the winning times for the years 1981 to 2000. Use two lines per stem.
(c) Compare the two distributions. How many times under 15 minutes are in each distribution?

6. *Split Stem: Golf* The U.S. Open Golf Tournament was played at Congressional Country Club, Bethesda, Maryland, with prizes ranging from $465,000 for first place to $5000. Par for the course is 70. The tournament consists of four rounds played on different days. The scores for each round of the 32 players who placed in the money (more than $17,000) were given on a web site. For more information, visit the Brase/Brase statistics site at **http://math.college.hmco.com/students** and find the link to golf. The scores for the first round were as follows:

71	65	67	73	74	73	71	71	74	73	71
70	75	71	72	71	75	75	71	71	74	75
66	75	75	75	71	72	72	73	71	67	

The scores for the fourth round for these players were as follows:

69	69	73	74	72	72	70	71	71	70	72
73	73	72	71	71	71	69	70	71	72	73
74	72	71	68	69	70	69	71	73	74	

(a) Make a stem-and-leaf display for the first-round scores. Use two lines per stem. (See Problem 5.)
(b) Make a stem-and-leaf display for the fourth-round scores. Use two lines per stem.
(c) Compare the two distributions. How do the highest scores compare? How do the lowest scores compare?

Are cigarettes bad for people? Cigarette smoking involves tar, carbon monoxide, and nicotine. The first two are definitely not good for a person's health, and the last ingredient can cause addiction. Problems 7, 8, and 9 refer to Table 2-15, which was taken from the web site maintained by the *Journal of Statistics Education*. For more information, visit the Brase/Brase statistics site at **http://math.college.hmco.com/students** and find the link to the *Journal of Statistics Education*. Follow the links to the cigarette data.

TABLE 2-15 Milligrams of Tar, Nicotine, and Carbon Monoxide (CO) per One Cigarette

Brand	Tar	Nicotine	CO	Brand	Tar	Nicotine	CO
Alpine	14.1	0.86	13.6	MultiFilter	11.4	0.78	10.2
Benson & Hedges	16.0	1.06	16.6	Newport Lights	9.0	0.74	9.5
Bull Durham	29.8	2.03	23.5	Now	1.0	0.13	1.5
Camel Lights	8.0	0.67	10.2	Old Gold	17.0	1.26	18.5
Carlton	4.1	0.40	5.4	Pall Mall Lights	12.8	1.08	12.6
Chesterfield	15.0	1.04	15.0	Raleigh	15.8	0.96	17.5
Golden Lights	8.8	0.76	9.0	Salem Ultra	4.5	0.42	4.9
Kent	12.4	0.95	12.3	Tareyton	14.5	1.01	15.9
Kool	16.6	1.12	16.3	True	7.3	0.61	8.5
L&M	14.9	1.02	15.4	Viceroy Rich Light	8.6	0.69	10.6
Lark Lights	13.7	1.01	13.0	Virginia Slim	15.2	1.02	13.9
Marlboro	15.1	0.90	14.4	Winston Lights	12.0	0.82	14.9
Merit	7.8	0.57	10.0				

Source: Journal of Statistics Education web site at http://www.amstat.org/publications/jse. Reprinted with permission.

7. *Health: Cigarette Smoke* Use the data in Table 2-15 to make a stem-and-leaf display for milligrams of tar per cigarette smoked.

8. *Health: Cigarette Smoke* Use the data in Table 2-15 to make a stem-and-leaf display for milligrams of carbon monoxide per cigarette smoked.

9. *Health: Cigarette Smoke* Use the data in Table 2-15 to make a stem-and-leaf display for milligrams of nicotine per cigarette smoked. In this case, truncate the measurements at the tenths position and use two lines per stem (see Problem 5, part a).

10. *Expand Your Knowledge: Back-to-Back Stem Plot* In archaeology, the depth (below surface grade) at which artifacts are found is very important. Greater depths sometimes indicate older artifacts, perhaps from a different archaeological period. Figure 2-17 is a *back-to-back stem plot* showing the depths of artifact locations at two different archaeological sites. These sites are from similar geographic locations. Notice that the stems are in the center of the diagram. The leaves for Site I artifact depths are shown to the left of the stem, while the leaves for Site II are to the right of the stem (see *Mimbres Mogollon Archaeology* by A. I. Woosley and A. J. McIntyre, University of New Mexico Press).
    (a) What are the least and greatest depths of artifact finds at Site I? at Site II?
    (b) Describe the data distribution of depths of artifact finds at Site I and at Site II.
    (c) At Site II, there is a gap in the depths at which artifacts were found. Does the Site II data distribution suggest that there might have been a period of no occupation?

FIGURE 2-17

Depth (in cm) of Artifact Location

5 | 2 | 0 = 25 cm at Site I and 20 cm at Site II

Site I		Site II
5	2	0 5 5
5 0	3	0 0 0 0 5 5
5 5 5 5	4	0 0 0 5
5 0	5	0 0 5 5
5 5 5 5 5 0	6	0 0 0 5 5 5
5 5 5 5 5 5 0 0	7	
5 5 0 0 0	8	
5 5 0 0	9	
5 5	10	
0	11	0 0 5 5 5 5
	12	0 0 0 0 5

SUMMARY

Organizing and presenting data are the main purposes of the branch of statistics called descriptive statistics. In this chapter, we have studied histograms, relative-frequency histograms, bar graphs, Pareto charts, circle graphs, time-series graphs, and stem-and-leaf displays. From the viewpoint of future applications, histograms are the most important because the area under a bar can represent the likelihood of data values falling into that class. Histograms and stem-and-leaf displays both reveal distribution properties such as uniformity, symmetry, or skewness.

IMPORTANT WORDS & SYMBOLS

Section 2.1

Frequency
Frequency distribution
Class width
Class, lower limit, upper limit
Class frequency
Class midpoint
Class mark
Frequency table
Class boundaries
Histogram
Relative-frequency table
Relative-frequency histogram
Symmetric distribution
Uniform distribution
Skewed left
Skewed right
Bimodal distribution
Dotplot

Section 2.2

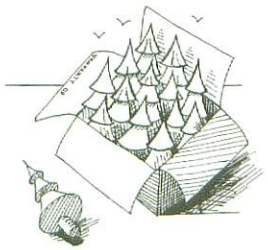
Bar graph
Pareto chart
Pie chart or circle graph
Time-series graph
Time series

Section 2.3

EDA
Stem
Leaf
Stem-and-leaf display
Back-to-back stem plot

VIEWPOINT

"This land is your land, This land is my land"*
—Woody Guthrie



But who actually owns the forest? On many maps, forest land (including national forests) is colored green. Such maps give the impression that vast areas of the western United States are public land. This is not the case! *USA Today* gave the following information about ownership of U.S. timber lands: state/local, 17%; federal, 10%; forest industry, 14%; and private nonindustry, 59%. Organize these data for better visual presentation using a Pareto chart and a circle graph.

*Words and music by Woody Guthrie TRO © Copyright 1956 (Renewed) 1958 (Renewed) Ludlow Music, Inc., New York, New York. Used by permission.

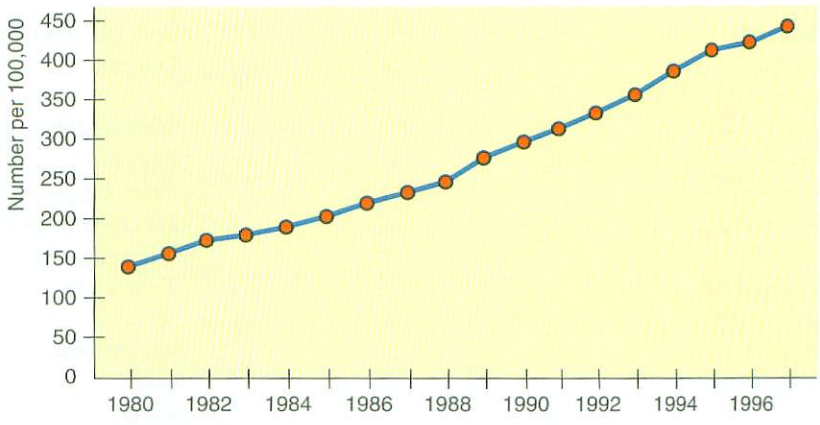
CHAPTER REVIEW PROBLEMS

1. *Focus Problem: Fuel Economy* Solve the focus problem at the beginning of this chapter.

2. *Criminal Justice: Prisoners* The time plot in Figure 2-18 on the next page gives the number of state and federal prisoners per 100,000 population (Source: *Statistical Abstract of the United States*, 120th Edition).
   (a) Estimate the number of prisoners per 100,000 people for 1980 and for 1997.
   (b) During the time period shown there was increased prosecution of drug offenses, longer sentences for common crimes, and reduced access to parole. What does

FIGURE 2-18

Number of State and Federal Prisoners per 100,000 Population



the time-series graph say about the prison population change per 100,000 people?

(c) In 1997, the U.S. population was approximately 266,574,000 people. At the rate of 444 prisoners per 100,000 population, about how many prisoners were in the system? The projected U.S. population for the year 2020 is 323,724,000. If the rate of prisoners per 100,000 stays the same as in 1997, about how many prisoners do you expect will be in the system in 2020? To obtain the most recent information, visit the Brase/Brase statistics site at **http://math.college.hmco.com/students** and find the link to the Census Bureau.

3. *IRS: Tax Returns* Almost everyone files (or will sometime file) a federal income tax return. A research poll for Turbo Tax (a computer software package to aid in tax-return preparation) asked what aspect of filing a return people thought to be the most difficult. The results showed that 43% of the respondents said understanding the IRS jargon, 28% said knowing deductions, 10% said getting the right form, 8% said calculating the numbers, and 10% didn't know. Make a circle graph to display this information. *Note:* Percentages will not total 100% because of rounding.

4. *Law Enforcement: DUI* Driving under the influence of alcohol (DUI) is a serious offense. The following data give the ages of a random sample of 50 drivers arrested while driving under the influence of alcohol. This distribution is based on the age distribution of DUI arrests given in the *Statistical Abstract of the United States* (112th Edition).

46	16	41	26	22	33	30	22	36	34
63	21	26	18	27	24	31	38	26	55
31	47	27	43	35	22	64	40	58	20
49	37	53	25	29	32	23	49	39	40
24	56	30	51	21	45	27	34	47	35

(a) Make a stem-and-leaf display of the age distribution.
(b) Make a frequency table using seven classes.
(c) Make a histogram showing class boundaries.

5. *Agriculture: Apple Trees* The following data represent trunk circumferences (in mm) for a random sample of 60 four-year-old apple trees at East Malling Agriculture Research Station in England (Reference: S. C. Pearce, University of Kent at Canterbury). *Note:* These data are also available with other software on the statSpace CD-ROM.

108	99	106	102	115	120	120	117	122	142
106	111	119	109	125	108	116	105	117	123
103	114	101	99	112	120	108	91	115	109
114	105	99	122	106	113	114	75	96	124
91	102	108	110	83	90	69	117	84	142
122	113	105	112	117	122	129	100	138	117

(a) Make a frequency table with seven classes showing class limits, class boundaries, midpoints, frequencies, and relative frequencies.
(b) Draw a histogram.
(c) Draw a relative-frequency histogram.

6. *Law: Corporation Lawsuits* Many people say the civil justice system is overburdened. Many cases center on suits involving businesses. The following data are based on a *Wall Street Journal* report. Researchers conducted a study of lawsuits involving 1908 businesses ranked in the Fortune 1000 over a 20-year period. They found the following distribution of civil justice caseloads brought before the federal courts involving the businesses:

Case Type	Number of Filings (in thousands)
Contracts	107
General torts (personal injury)	191
Asbestos liability	49
Other product liability	38
All other	21

Note: Contracts cases involve disputes over contracts between businesses.
(a) Make a Pareto chart of the caseloads. Which type of cases occur most frequently?
(b) Make a circle chart showing the percentage of cases of each type.

7. *Archaeology: Tree-Ring Data* *The Sand Canyon Archaeological Project,* edited by W. D. Lipe and published by Crow Canyon Archaeological Center, contains the stem-and-leaf diagram in Figure 2-19 on the next page. The study uses tree rings to accurately determine the year in which a tree was cut. The figure gives the tree-ring-cutting dates for samples of timbers found in the architectural units at Sand Canyon Pueblo. The text referring to the figure says, "The three-digit numbers in the left column represent centuries and decades A.D. The numbers to the right represent individual years, with each number derived from an individual sample. Thus, **124 2 2 2** represents three samples dated to A.D. 1242." Use Figure 2-19 and the verbal description to answer the following questions.
(a) Which decade contained the most samples?
(b) How many samples had a tree-ring-cutting date between *1200* A.D. and *1239* A.D., inclusive?

FIGURE 2-19

Tree-Ring-Cutting Dates from Architectural Units at Sand Canyon Pueblo: *The Sand Canyon Archaeological Project*

119	56
120	00123333333333333333333333
120	
121	2
121	55
122	0011112234444444
122	589
123	012334
123	5555555555555556889
124	12222222222222222222222344
124	5689999999999
125	00000000000000011111111222
125	
126	00012222222222224444444
126	555667
127	0111144

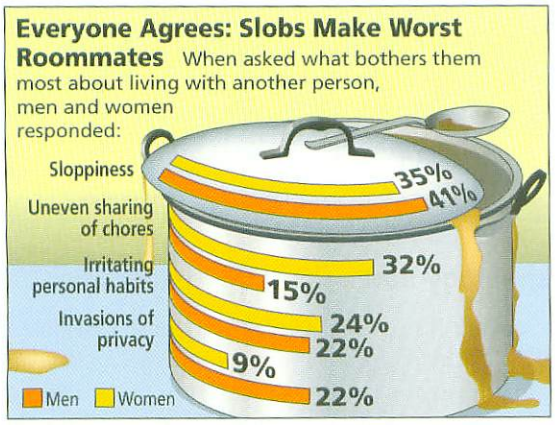
(c) What are the dates of the longest interval during which no tree-cutting samples occurred? What might this indicate about new construction or renovation of the pueblo structures during this period?

DATA HIGHLIGHTS: GROUP PROJECTS

Break into small groups and discuss the following topics. Organize a brief outline in which you summarize the main points of your group discussion.

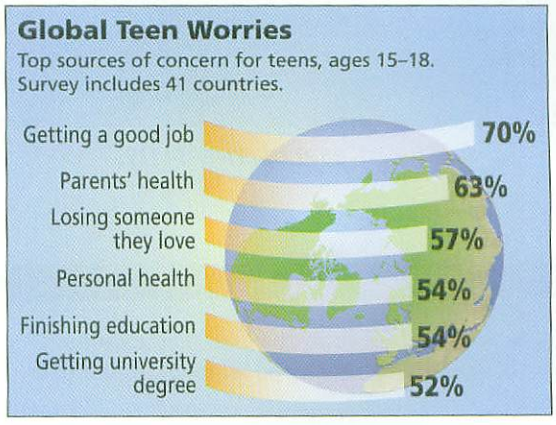
1. Examine Figure 2-20, "Everyone Agrees: Slobs Make Worst Roommates." This is a double bar graph because two percentages are given for each response category: responses from men and responses from women. Comment about how the artistic rendition has slightly changed the format of a bar graph. Do the bars seem to have lengths that accurately reflect the relative percentages of the responses? In your own opinion, does the artistic rendition enhance or confuse the information? Explain. Which characteristic of "worst roommates" does the graphic seem to illustrate? Can this graph be considered a Pareto chart for men? for women? Why or why not? From the information given in the figure, do you think the survey just listed the four given annoying characteristics? Do you think a respondent could choose more than one characteristic? Explain your answer in terms of the percentages given and in terms of the explanation given in the graphic. Could this information also be displayed in one circle graph for men and another for women? Explain.
2. Examine Figure 2-21, "Global Teen Worries." How many countries were contained in the sample? The graph contains bars and a circle. Which bar is the longest? Which bar represents the greatest percentage? Is this a bar graph or not? If not, what changes would need to be made to put the information into a bar graph? Could the graph be made into a Pareto chart? Could it be made into a circle graph? Explain.

FIGURE 2-20



Source: Advantage Business Research for Mattel *Compatibility*

FIGURE 2-21



Source: BrainWaves Group's New World Teen Study

LINKING CONCEPTS: WRITING PROJECTS

Discuss each of the following topics in class or review the topics on your own. Then write a brief but complete essay in which you summarize the main points. Please include formulas and graphs as appropriate.

1. In your own words, explain the differences among histograms, relative-frequency histograms, bar graphs, circle graphs, time-series graphs, Pareto charts, and stem-and-leaf displays. If you have nominal data, which graphic displays might be useful? What if you have ordinal, interval, or ratio data?

2. What do we mean when we say a histogram is skewed to the left? to the right? What is a bimodal histogram? Discuss the following statement: "A bimodal histogram usually results if we draw a sample from two populations at once." Suppose you took a sample of weights of college football players and with this sample you included weights of cheerleaders. Do you think a histogram made from the combined weights would be bimodal? Explain.

3. Discuss the statement that stem-and-leaf displays are quick and easy to construct. How can we use a stem-and-leaf display to make the construction of a frequency table easier? How does a stem-and-leaf display help you spot extreme values quickly?

4. Go to the library and pick up a current issue of *The Wall Street Journal*, *Newsweek*, *Time*, *USA Today*, or other news media. Examine each newspaper or magazine for graphs of the type discussed in this chapter. List the variables used, method of data collection, and general type of conclusion drawn from the graphs. Another source for information is the Internet. Explore several web sites, and categorize the graphs you find as you did for the print media. For interesting web sites, visit the Brase/Brase statistics site at **http://math.college.hmco.com/students** and find links to the Social Statistics Briefing Room, to law enforcement, and to golf.

Using Technology

TI-84PLUS/TI-83PLUS • EXCEL • MINITAB • SPSS

APPLICATIONS

The following tables show the first-round winning scores of the NCAA men's and women's basketball teams.

TABLE 2-16 Men's Winning First-Round NCAA Tournament Scores

95	70	79	99	83	72	79	101
69	82	86	70	79	69	69	70
95	70	77	61	69	68	69	72
89	66	84	77	50	83	63	58

TABLE 2-17 Women's Winning First-Round NCAA Tournament Scores

80	68	51	80	83	75	77	100
96	68	89	80	67	84	76	70
98	81	79	89	98	83	72	100
101	83	66	76	77	84	71	77

1. Use the software or method of your choice to construct separate histograms for the men's and women's winning scores. Try 5, 7, and 10 classes for each. Which number of classes seems to be the best choice? Why?
2. Use the same class boundaries for histograms of men's and of women's scores. How do the scores for the two groups compare? What general shape do the histograms follow?
3. Use the software or method of your choice to make stem-and-leaf displays for each set of scores. If your software does not make stem-and-leaf displays, sort the data first and then make a back-to-back display by hand. Do there seem to be any extreme values in either set? How do the data sets compare?

Technology Hints: Creating Histograms

The default histograms produced by the TI-84Plus/TI-83Plus calculators, Minitab, and Excel all determine automatically the number of classes to use. To control the number of classes the technology uses, follow the key steps as indicated. The display screens are generated for data found in Table 2-1—Commuting Distances of Dallas Workers.

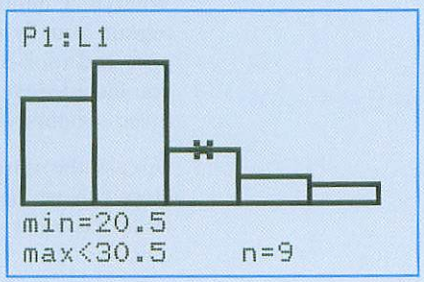
TI-84Plus/TI-83Plus

Determine the class width for the number of classes you want and the lower class boundary for the first class. Enter the data in list L1.

Press **STATPLOT** and highlight On and the histogram plot.

Press **WINDOW** and set Xmin = lowest class boundary, Xscl = class width. Use appropriate values for the other settings.

Press **GRAPH. TRACE** gives boundaries and frequency.

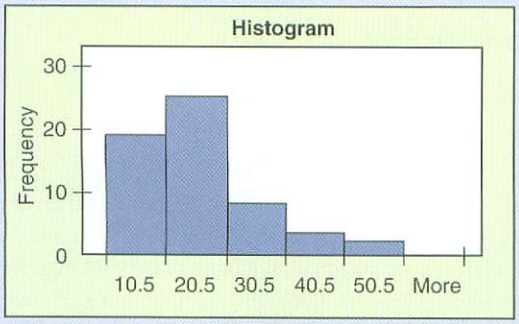


Excel

Determine the upper class boundaries for the five classes. Enter the data. In a separate column, enter the upper class boundaries. Use the menu selection **Tools ➤ Data Analysis ➤ Histogram.**

Put the data range in the Input Range. Put the upper class boundaries range in the Bin Range.

To make bars touch, right click on a bar and select **Format Data Series ➤ Options tab.** Set the **gap width** to 0.

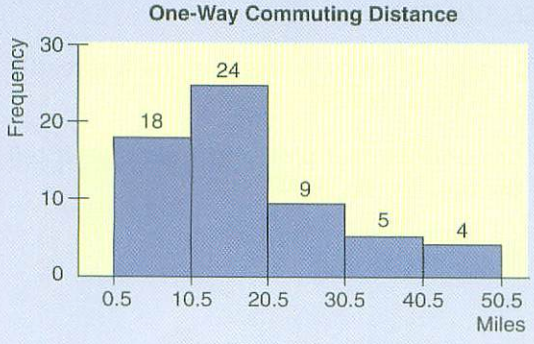


Minitab

Determine the class boundaries. Enter the data. Use the menu selection **Graph ➤ Histogram.**

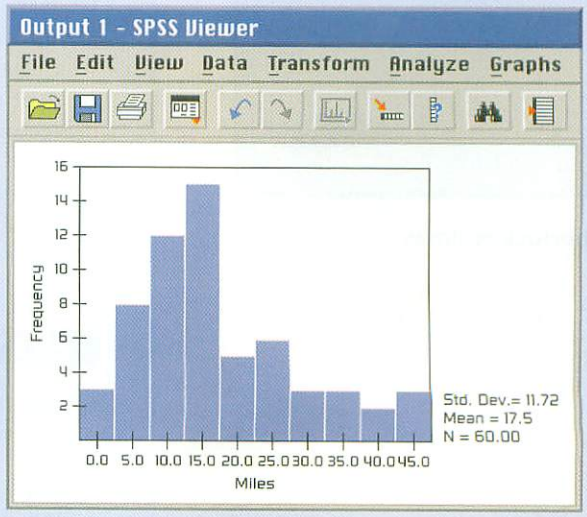
In the Dialogue Box, press **Options.** Then select cutpoints and enter the class boundaries as cutpoint positions.

Select **Frame** to adjust scales. Otherwise, press **OK.**



SPSS

The SPSS screenshot shows the default histogram created by the menu choices **Analyze ➤ Descriptive Statistics ➤ Frequencies.** In the dialogue box, move the variable containing the data into the variables window. Click **Charts** and select **Histograms.** Click the Continue button and then the OK button. In SPSS version 12, there are procedures to control the boundaries (cutpoints) of the histogram.



Specific instructions for setting class boundaries (cut points) of a histogram are provided in the Technology Guide that accompanies this text.