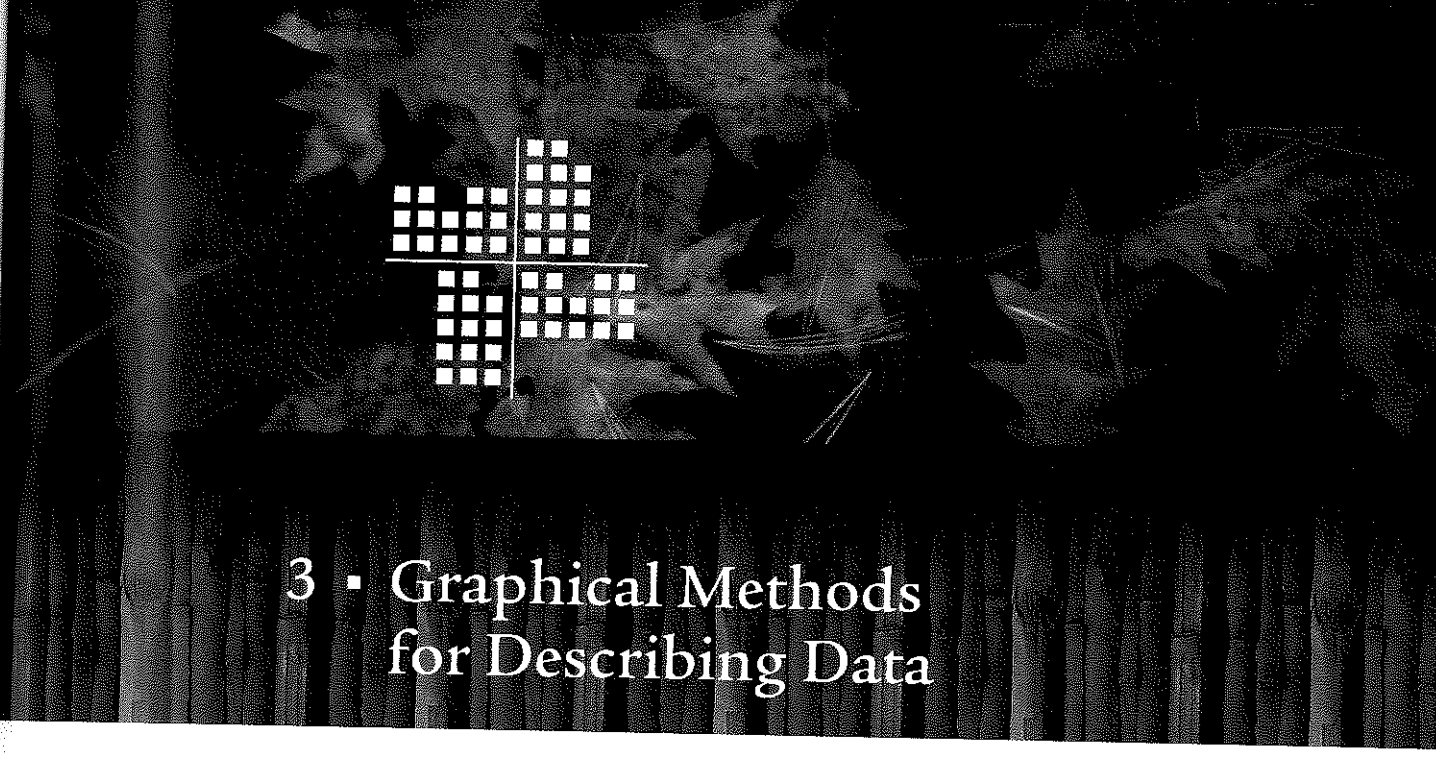


3 • Graphical Methods for Describing Data

StatisticsNow™

Throughout the chapter, this icon introduces a list of resources on the StatisticsNow website at **http://1pass.thomson.com** that will:

- Help you evaluate your knowledge of the material
- Allow you to take an exam-prep quiz
- Provide a Personalized Learning Plan targeting resources that address areas you should study

Most college students (and their parents) are concerned about the cost of a college education. *The Chronicle of Higher Education* (August 31, 2001) reported the average tuition and fees for 4-year public institutions in each of the 50 U.S. states for the 1999–2000 academic year. The accompanying values are given in alphabetical order by state:

2833	2855	2252	2785	2559	2775	4435	4642	2244	2524
2965	2458	4038	3646	2998	2439	2723	2430	4122	4552
4105	4538	3800	2872	3701	3011	2930	2034	6083	5255
2340	3983	2054	2990	4495	2183	3582	5610	4318	3638
3210	2698	2644	2147	6913	3733	3357	2549	3313	2416

A number of interesting questions could be posed about these data. What is a typical or representative value of average tuition and fees for the 50 states? Are observations concentrated near the typical value, or does the value of average tuition and fees differ quite a bit from state to state? Do states with low tuition and fees or states with high tuition and fees predominate, or are there roughly equal numbers of the two types? Are there any states whose average tuition and fees are somehow unusual compared to the rest of the data? What proportion or fraction of the states have average tuition and fees exceeding $3000? exceeding $5000?

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 Perceived Risk of Smoking

StatisticsNow™

Explore this example with technology. Download your technology manual at the StatisticsNow website or from your CD.

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.

	Relative Frequency		
Perceived Risk of Smoking	**Smokers**	**Former Smokers**	**Nonsmokers**
Very harmful	.60	.78	.86
Somewhat harmful	.30	.16	.10
Not too harmful	.07	.04	.03
Not harmful at all	.03	.02	.01

In constructing a comparative bar graph, we should use relative frequency to construct the scale on the vertical axis so that we can make meaningful comparisons when 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 among the three groups with respect to the perceived risk of smoking. The proportion believing that smoking is very harmful is noticeably smaller for smokers than for either former smokers or nonsmokers, and the proportion for former smokers is smaller than the proportion for nonsmokers.

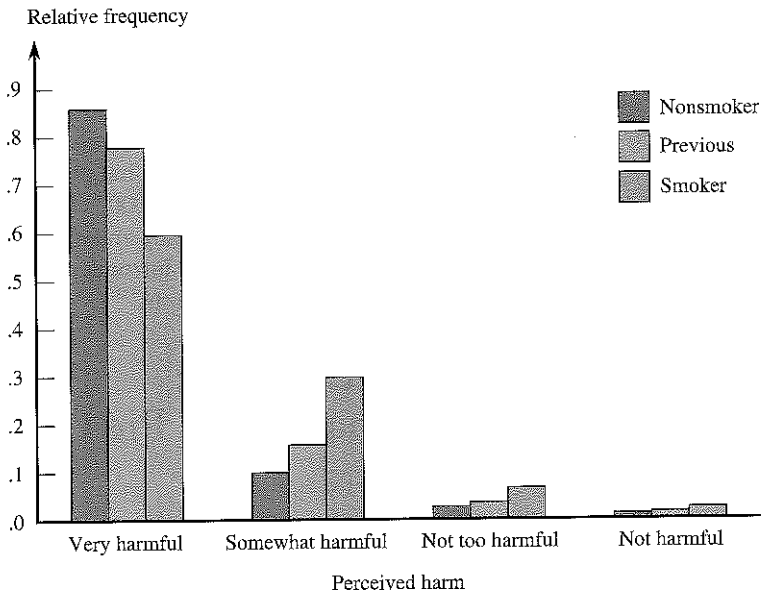


Figure 3.1 Comparative bar chart of perceived harm of smoking.

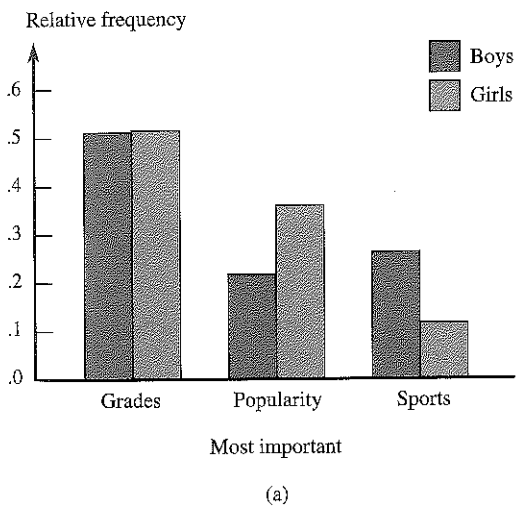
▪ Example 3.2 What Could Be More Important Than Being Popular?

To see why it is important to use relative frequencies rather than frequencies to compare groups of different sizes, let's look at data from a survey of 227 boys and 251 girls in grades 4 through 6. Each student was asked what he or she thought was most important: getting good grades, being popular, or being good at sports. The resulting data, from the paper "The Role of Sport as a Social Determinant for Children" (*Research Quarterly for Exercise and Sport* [1992]: 418–424), are summarized in the accompanying table:

	Boys		Girls	
Most Important	Frequency	Relative Frequency	Frequency	Relative Frequency
Grades	117	.515	130	.518
Popularity	50	.220	91	.363
Sports	60	.264	30	.120
Total	**227**	**.999**	**251**	**1.001**

(Note: The relative frequencies for each group should sum to 1; here the small discrepancy is due to rounding in the relative frequencies.)

Figure 3.2(a) shows a correctly constructed comparative bar chart based on relative frequencies. Figure 3.2(b) shows an *incorrect* comparative bar chart constructed using frequencies rather than relative frequencies to determine the height of each bar. Notice that the graph of Figure 3.2(b) is misleading in that it does not accurately convey the differences between the two groups.



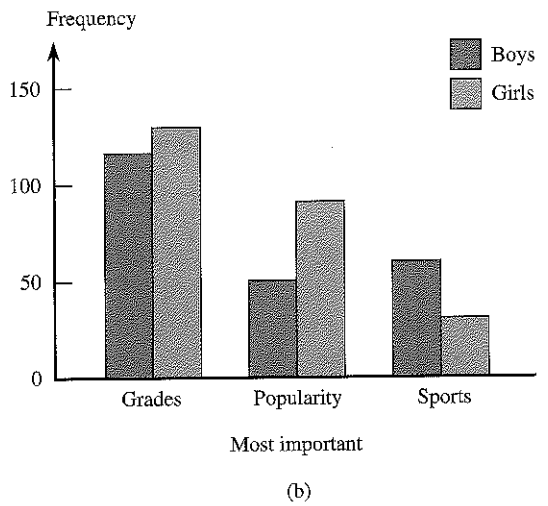


Figure 3.2 Correctly and incorrectly constructed comparative bar charts: (a) correctly constructed comparative bar chart (based on relative frequencies); (b) incorrectly constructed comparative bar chart (based on frequencies).

As you can see from Example 3.2, comparative bar charts should be constructed using relative frequencies. It is reasonable to use frequencies only if the number of observations is the same for *all* the groups being compared.

▪ 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.

▪ Example 3.3 College Choice

The Chronicle of Higher Education (August 31, 2001) published data collected in a survey of a large number of students who were college freshmen in the fall of 2001. One question asked whether the student was attending his or her first, second, or third choice of university. Fourth or higher choices were combined in a category called "other." The resulting data are summarized in the pie chart of Figure 3.3.

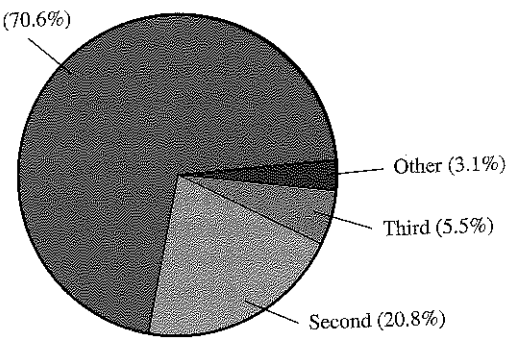


Figure 3.3 Pie chart of data on college choice.

Pie charts are most effective for summarizing data sets when there are not too many different categories.

▪ Pie Chart for Categorical Data

When to Use

Categorical data with a relatively small number of possible categories. Sometimes this is achieved by using an "other" category. 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:

$$\text{slice size} = 360(\text{category relative frequency})$$

(because there are 360 degrees in a circle).

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 3.4 Birds That Fish

StatisticsNow™

Explore this example with technology. Download your technology manual at the StatisticsNow website or from your CD.

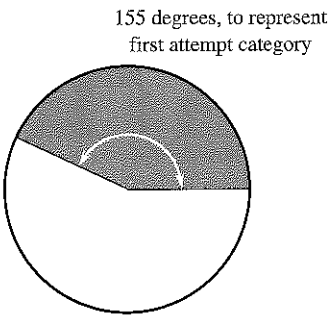
Night herons and cattle egrets are species of birds that feed on aquatic prey in shallow water. These birds wade through shallow water, stalking submerged prey and then striking rapidly and downward through the water in an attempt to catch the prey. The article "Cattle Egrets Are Less Able to Cope with Light Refraction Than Are Other Herons" (*Animal Behaviour* [1999]: 687–694) gave data on outcome when 240 cattle egrets attempted to capture submerged prey. The data are summarized in the following frequency distribution:

Outcome	Frequency	Relative Frequency
Prey caught on first attempt	103	.43
Prey caught on second attempt	41	.17
Prey caught on third attempt	2	.01
Prey not caught	94	.39

To draw a pie chart by hand, we must first compute the slice size for each category. This is done as follows:

Category	Slice Size
First attempt	(.43)(360) = 154.8°
Second attempt	(.17)(360) = 61.2°
Third attempt	(.01)(360) = 3.6°
Not caught	(.39)(360) = 140.4°

We would then draw a circle and use a protractor to mark off a slice corresponding to about 155°, as illustrated here:



Continuing to add slices in this way leads to a completed pie chart.

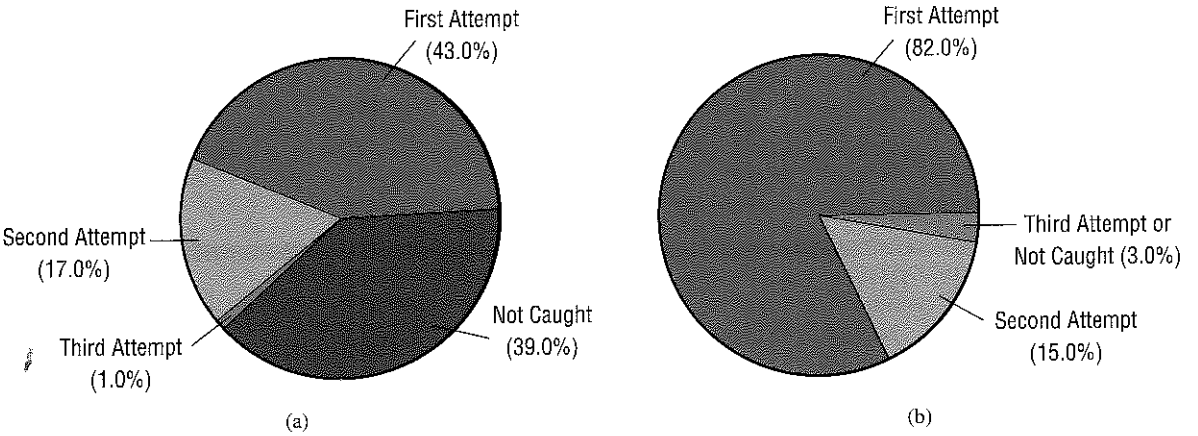
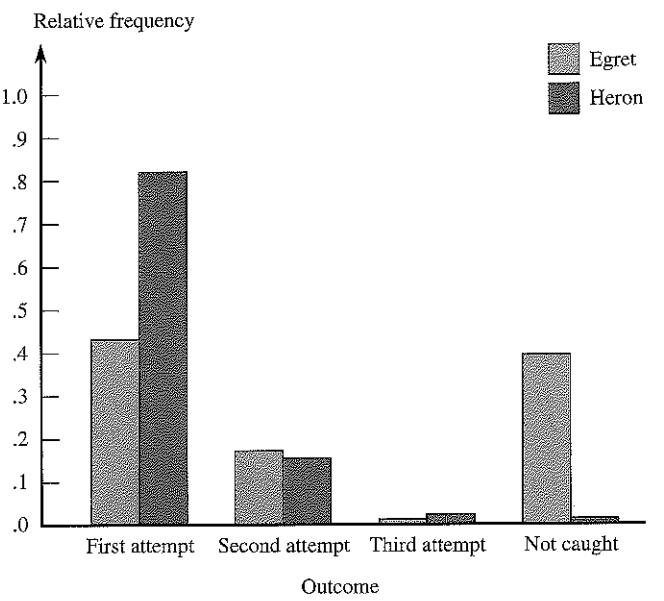


Figure 3.4 Pie charts for Example 3.4: (a) cattle egret data; (b) night heron data.

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 cattle egret data was created with MINITAB and is shown in Figure 3.4(a). Figure 3.4(b) shows a pie chart constructed using similar data on outcome for 180 night herons. Although some differences between night herons and cattle egrets can be seen by comparing the pie charts in Figures 3.4(a) and 3.4(b), it is difficult to actually compare category proportions using pie charts. A comparative bar chart (Figure 3.5) makes this type of comparison easier.

Figure 3.5 Comparative bar chart for the egret and heron data.



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

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 chart** (also sometimes called a stacked bar chart) avoids these difficulties by using a rectangular bar rather than a circle to represent the entire data set. The bar is then divided into segments, with different segments representing different categories. As with pie charts, the size 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 Where Are the Men?

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. To construct a segmented bar graph for these data, first draw a bar of any fixed width and length. Then divide the bar into two segments, one for males and one for females. The length of the segment for males is (.42)(length of the bar) because the relative frequency for this category is .42. The segmented bar can be displayed either vertically or horizontally, as shown in Figure 3.6.

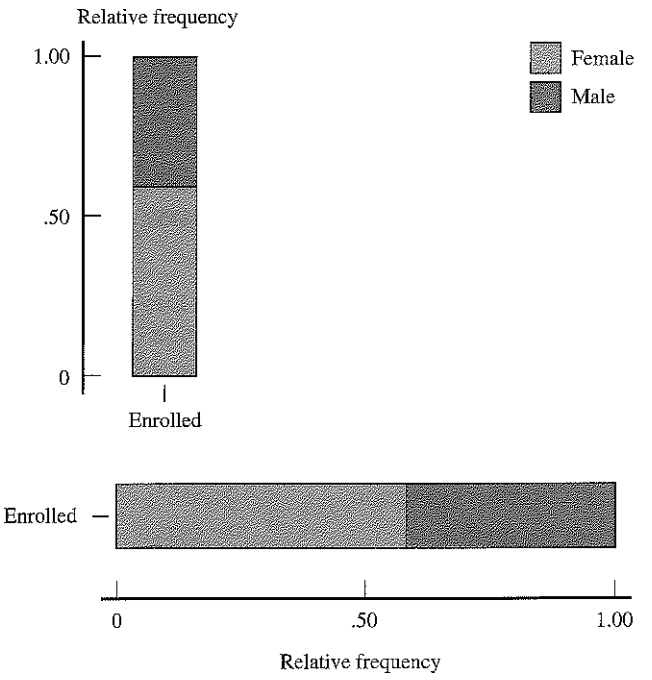


Figure 3.6 Segmented bar graphs for the gender data of Example 3.5.

Segmented bar graphs can also be used to compare different populations on the basis of a categorical variable, as illustrated in Example 3.6.

▪ Example 3.6 Sex on TV

The segmented bar graph shown in Figure 3.7 appeared in the article "Study: More TV Shows Depicting Sexuality" (Associated Press, February 7, 2001). Based on this graph, it is easy to see how the proportion of television shows with sexual content differs for the different types of television shows studied.

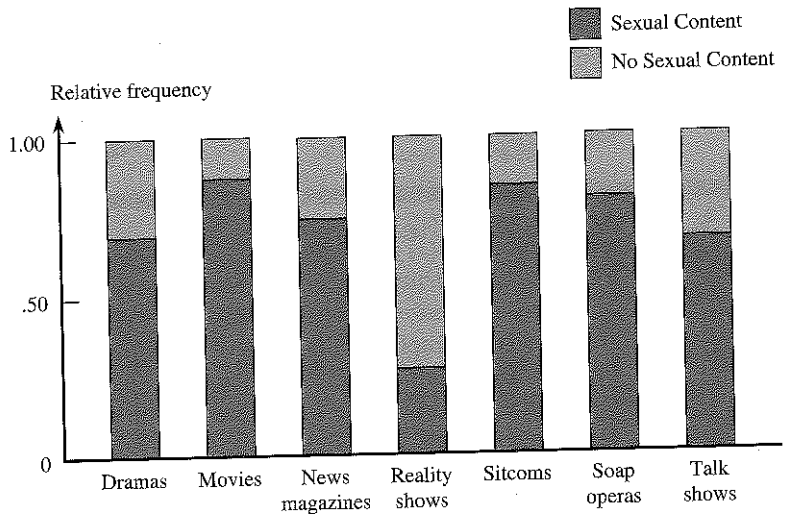


Figure 3.7 Segmented bar graph for the data of Example 3.6.

▪ 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 Example 3.7.

▪ Example 3.7 Wine Grape Production

The 1998 Grape Crush Report for San Luis Obispo and Santa Barbara counties in California gave the following information on grape production for each of seven different types of grapes used to make wine (*San Luis Obispo Tribune*, February 12, 1999):

Type of Grape	Tons Produced
Cabernet sauvignon	21,656
Chardonnay	39,582
Chenin blanc	1,601
Merlot	11,210
Pinot noir	2,856
Sauvignon blanc	5,868
Zinfandel	7,330
Total	**90,103**

Although this table is not a frequency distribution for a categorical data set, it is common to represent information of this type graphically using either a pie chart or a bar chart. A pie chart is shown in Figure 3.8(a). The pie represents the total grape production, and the slices show the proportion of the total production for each of the seven types of grapes. Figure 3.8(b) shows a bar chart representation of the grape production data.

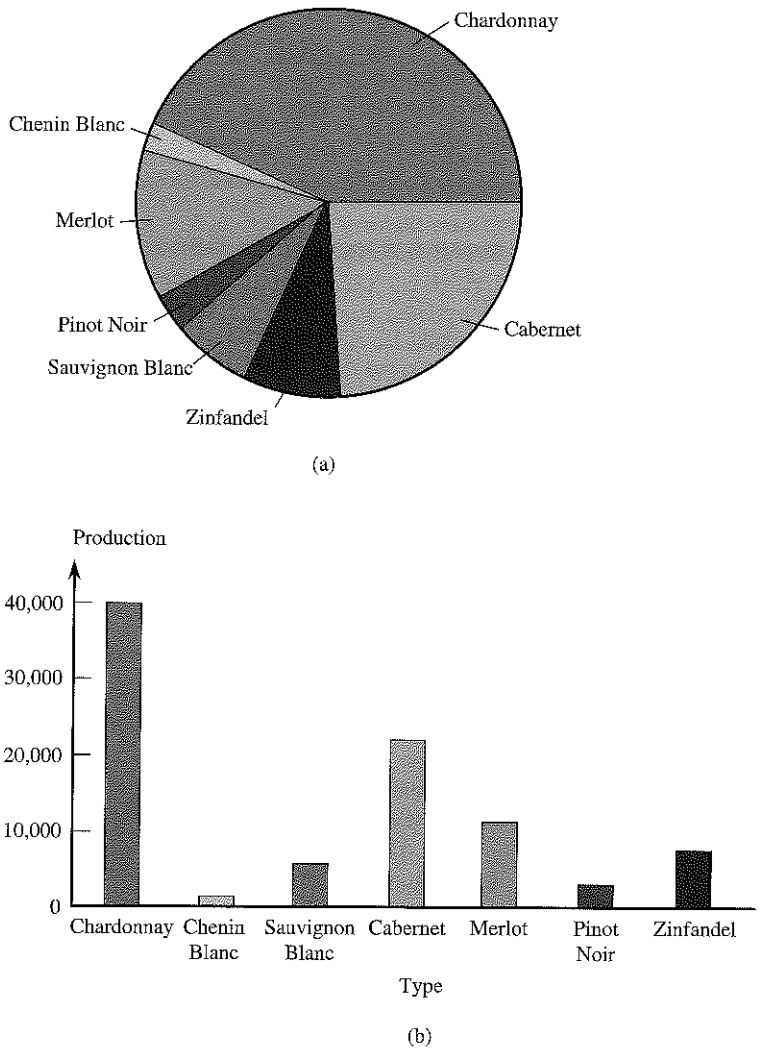


Figure 3.8 Grape production data: (a) pie chart; (b) bar chart.

The article also gave grape production information for 1997 and commented on the effect of harsh weather on the counties' grape crop. Figure 3.9 shows a comparative bar chart of grape production in 1997 and 1998. The comparative bar chart clearly shows decreased production in 1998 for all but two of the seven types of grape.

b. Summarize the given data using a segmented bar chart.

3.8 The paper "Exploring the Factors that Influence Men and Women to Form Medical Career Aspirations" (*Journal of College Student Development* [1998]: 417–421) gave estimates of the percentage of college freshmen nationwide who hoped to have a medical career. These percentages were based on an annual survey of "a nationally representative sample of all entering college students":

	Percentage with Medical Career Aspirations			
	1975	1980	1985	1990
Men	3.9	4.4	4.1	3.9
Women	2.5	2.9	3.4	3.7

Construct a comparative bar graph that shows the proportion with medical career aspirations for men and women over time. (Hint: Mark the years on the horizontal axis, and then construct a bar for each gender for each year.) Comment on any interesting features of the graph.

3.9 The paper referenced in Exercise 3.8 also gave information on parent's occupation for those who aspire to a medical career. The data are given in the following table:

Parent Occupation	Male	Female
Only mother is a physician	596	712
Only father is a physician	504	350
Both parents are physicians	734	824
Neither parent is a physician	166	114

Construct two pie charts, one for men and one for women, to display this information, and comment on the similarities and differences between the two.

3.10 The article "Physicians Accessing the Internet: The PAI Project" (*Journal of the American Medical Association* [1999]: 633–634) gave the following information on Internet usage by doctors. The data are from a survey of 324 randomly selected doctors who were asked to indicate the category that best described how often they used the Internet.

Internet Usage Pattern	Frequency
Never	31
Rarely (about 3 times per year)	15
Occasionally (about once a month)	52
Often (about once a week)	109
Daily	117

Construct a pie chart for these data, and write a brief summary that describes Internet usage for physicians who participated in this survey.

3.11 The paper "Sexual Content of Top-Grossing Motion Pictures" (*Journal of Health Education* [1998]: 354–357) gave the accompanying information on the ratings of the 10 movies that made the most money in the years 1987 and 1992:

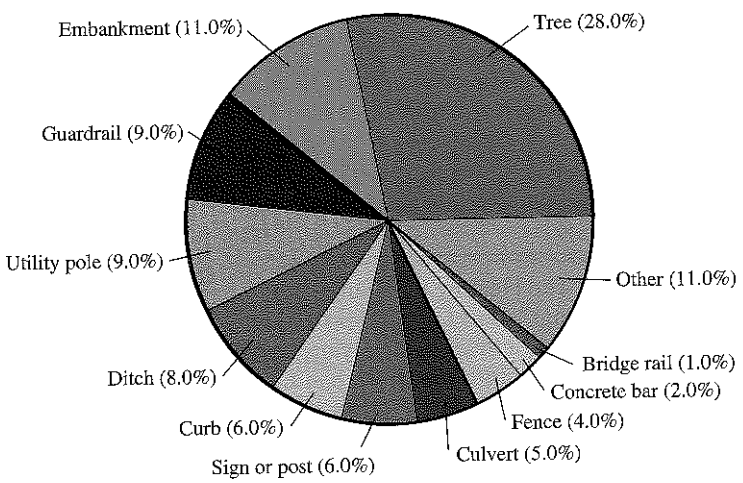
	Rating		
	G	PG/PG-13	R
1987	0	3	7
1992	1	5	4

a. Construct a pie chart to show the distribution of ratings for 1987.

b. Construct a pie chart to show the distribution of ratings for 1992, and comment on the differences between this chart and the one based on the 1987 data.

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

Pie chart for Exercise 3.12



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.

3.13 The percentage of U.S. gross domestic product (GDP) spent on health care over the period 1960–1995 was given in the paper "Building the Next Generation of Healthy People" (*Public Health Reports* [1999]: 213–216). Construct a bar chart for the data given in the accompanying table, and comment on the interesting features of this graph. Be sure to comment on the trend over time.

Year	Percentage of GDP Spent on Health Care
1960	15.1
1965	15.7
1970	17.1
1975	18.0
1980	18.9
1985	10.3
1990	12.2
1995	13.7

3.14 Bizrate.com reported the accompanying data on Internet sales (*San Luis Obispo Tribune*, April 26, 2002):

Year	Internet Sales (billions of dollars)
2000	28.9
2001	35.9
2002	51.5

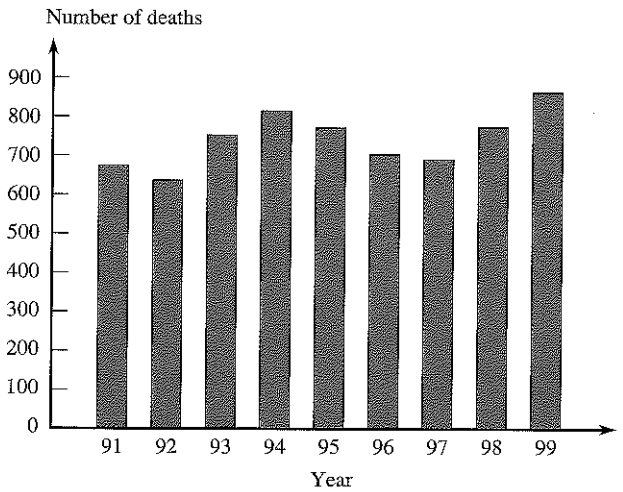
Construct a graphical display that illustrates the trend over time. (Hint: Would a pie chart or a bar chart be the most appropriate display?)

3.15 The source referenced in Exercise 3.14 also gave data on sales (in billions of dollars) for 2002 in major retailing categories, as shown:

Category	Sales
Travel	7.0
Computer hardware	2.4
Office supplies	1.7
Apparel and accessories	1.3
Consumer electronics	0.744
Event tickets	0.581
Books	0.557
Home and garden	0.458
Health and beauty	0.308
Sports and fitness	0.258
Computer software	0.236

Construct a graphical display to summarize the data in this table. Did you choose a bar chart or a pie chart for your display? Explain your choice.

3.16 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, one called the stem and the other called the 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 Binge Drinking

StatisticsNow™

Explore this example with technology. Download your technology manual at the StatisticsNow website or from your CD.

The use of alcohol by college students is of great concern, not only to those in the academic community but also, because of potential health and safety consequences, to society at large. The article "Health and Behavioral Consequences of Binge Drinking in College" (*Journal of the American Medical Association* [1994]: 1672–1677) reported on a comprehensive study of heavy drinking on campuses across the country. A binge episode was defined as five or more drinks in a row for males and four or more drinks in a row for females. Figure 3.10 shows a stem-and-leaf display of 140 values of the percentage of undergraduate students who are binge drinkers at various colleges. (These values were not given in the cited article, but our display agrees with a picture of the data that did appear.)

Figure 3.10 Stem-and-leaf display for percentage of binge drinkers at each of 140 colleges.

```
0 | 4
1 | 1345678889
2 | 1223456666777889999
3 | 011223334455566667777888899999
4 | 111222223344445566666677788888999
5 | 00111222233455666667777888899
6 | 01111244455666778
```

Stem: Tens digit
Leaf: Ones digit

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 digit

Leaf: Ones digit

tells us that the observation that had a stem of 2 and a leaf of 1 corresponds to a college where 21% (as opposed to 2.1% or 0.21%) of the students were binge drinkers.

The display in Figure 3.10 suggests that a typical or representative value is in the stem 4 row, perhaps someplace in the low 40% range. The observations are not highly concentrated about this typical value, as would be the case if all values were between 20% and 49%. The display rises to a single peak as we move downward and then declines, and there are no gaps in the display. The shape of the display is not perfectly symmetric but rather appears to stretch out a bit more in the direction of low stems than in the direction of high stems. The most surprising feature of these data is that at most colleges in the sample, at least one-quarter of the students are binge drinkers.

The leaves on each line of the display in Figure 3.10 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 in getting a sense of a typical value for the data set, as well as 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 data on binge drinking (Figure 3.10) does not show any outliers.

▪ Definition

An **outlier** is an unusually small or large data value. In Chapter 4 we give a precise rule for deciding when an observation is an outlier.

▪ Stem-and-Leaf Displays

When to Use

Numerical data sets with a small to moderate number of observations (does not work well with 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 such a value, the presence of any gaps in the data, the extent of symmetry in the distribution of values, the number and location of peaks, and the presence of any outliers.

▪ 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 2000 for the 50 U.S. states. The observations ranged from a low value of 2054 to a high value of 6913. The data are reproduced here:

2833	2855	2252	2785	2559	2775	4435	4642	2244	2524
2965	2458	4038	3646	2998	2439	2723	2430	4122	4552
4105	4538	3800	2872	3701	3011	2930	2034	6083	5255
2340	3983	2054	2990	4495	2183	3582	5610	4318	3638
3210	2698	2644	2147	6913	3733	3357	2549	3313	2416

A natural choice for the stem is the leading (thousands) digit. This would result in a display with 5 stems (2, 3, 4, 5, 6). Using the first two digits of a number as the

stem would result in 41 stems (20, 21, . . . , 60). A stem-and-leaf display with 41 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

$2833 \rightarrow$ stem = 2, leaf = 833

$2965 \rightarrow$ stem = 2, leaf = 965

$4105 \rightarrow$ stem = 4, leaf = 105

The leaves have been entered in the display of Figure 3.11 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 3.11 shows that most states had average tuition and fees in the $2000 range and that the typical average tuition and fees is around $3000. A few states have average tuition and fees at public four-year institutions that are quite a bit higher than most other states (the four states with the highest values were New Jersey, Pennsylvania, New Hampshire, and Vermont).

```
2 | 833,855,252,785,559,775,244,524,965,458,998,439,723,430,872,930,034,340,054,990,183,698,644,147,549,416
3 | 646,800,701,011,983,582,638,210,733,357,313
4 | 435,642,038,122,552,105,538,495,318
5 | 255,610
6 | 083,913
```

Stem: Thousands
Leaf: Ones

Figure 3.11 Stem-and-leaf display of average tuition and fees.

An alternative display (Figure 3.12) 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 a typical value, spread, or shape is lost in this truncation and the display is simpler and more compact.

```
2 | 8 8 2 7 5 7 2 5 9 4 9 4 7 4 8 9 0 3 0 9 1 6 6 1 5 4
3 | 6 8 7 0 9 5 6 2 7 3 3
4 | 4 6 0 1 5 1 5 4 3
5 | 2 6
6 | 0 9
```

Stem: Thousands
Leaf: Hundreds

Figure 3.12 Stem-and-leaf display of the average tuition and fees data using truncated leaves.

▪ 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 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 Protein Intake of Athletes

The accompanying data on daily protein intake (in grams of protein per kilogram of body weight) for 20 competitive athletes was obtained from a plot in the article "A Comparison of Plasma Glutamine Concentration in Athletes from Different Sports" (*Medicine and Science in Sports and Exercise* [1998]: 1693–1697):

1.4	2.2	2.7	1.5	2.3	1.7	2.3	1.5	1.8	2.8
1.8	1.9	2.0	2.3	1.5	1.9	1.7	1.8	1.6	3.0

Because each value in the data set has only two digits, we must use the first digit for the stem and the last digit for the leaf. The corresponding stem-and-leaf display is shown in Figure 3.13; the display has only three stems (1, 2, and 3), and all but one of the leaves are located at the 1 and 2 stems. A more informative display using repeated stems is shown in Figure 3.14.

Figure 3.13 Stem-and-leaf display for the protein intake data.

1	457588959786	
2	2733803	Stem: Ones
3	0	Leaf: Tenths

Figure 3.14 Stem-and-leaf display for protein intake data using repeated stems.

1L	4	
1H	57588959786	
2L	23303	
2H	78	Stem: Ones
3L	0	Leaf: Tenths

▪ 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 extend to the right of the stem values and the leaves for the second group extend to the left, can provide preliminary visual impressions and insights.

▪ Example 3.11 Tobacco Use in G-Rated Movies

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	9	2	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

To construct a stem-and-leaf display, think of each observation as a three-digit number. For example, 2 would be 002 and 37 would be 037. We then use the first digit of each number as the stem and the remaining two digits as the leaf. For simplicity, let's truncate the leaves to one digit:

223 → stem = 2, leaf = 2 (truncated from 23)

37 → 037 → stem = 0, leaf = 3 (truncated from 37)

9 → 009 → stem = 0, leaf = 0 (truncated from 09)

The resulting comparative stem-and-leaf display, using repeated stems, is shown in Figure 3.15.

Figure 3.15
Comparative stem-and-leaf display for tobacco exposure times.

Other Studios		Disney
12000	0L	33100020
59	0H	57
1	1L	
56	1H	756
All Dogs Go to Heaven 0	2L	20 *Pinocchio, James and the Giant Peach*
	2H	9 *101 Dalmatians*
	3L	
	3H	
	4L	
	4H	
	5L	4 *The Three Caballeros*

Stem: Hundreds
Leaf: Tens

One first impression is that there are many films where the tobacco exposure time is small, both for Disney films and for those made by the other studios. Disney has more films with longer exposure to tobacco, and there is an obvious outlier in the Disney data (the 548-sec exposure in the 1945 film *The Three Caballeros*).

▪ Exercises 3.17–3.24

3.17 The stem-and-leaf display at the top of page 97 shows observations on average shower flow rate (in liters per minute) for a sample of 129 houses in Perth, Australia ("An Application of Bayes Methodology to the Analysis of Diary Records from a Water Use Study," *Journal of the American Statistical Association* [1987]: 705–711).

a. What is the smallest flow rate in the sample?

b. If one additional house yielded a flow rate of 8.9, where would this observation be placed on the display?

c. What is a typical, or representative, flow rate?

d. Does the display appear to be highly concentrated, or quite spread out?

e. Does the distribution of values in the display appear to be reasonably symmetric? If not, how would you describe the departure from symmetry?

f. Does the data set appear to contain any outliers (observations far removed from the bulk of the data)?

3.18 The Connecticut Agricultural Experiment Station conducted a study of the calorie content of different types of beer. The calorie contents (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

2	23	
3	2344567789	
4	01356889	
5	00001114455666789	
6	000012222334445666778999	
7	00012233455555668	
8	02233448	
9	012233335666788	
10	2344455688	
11	2335999	
12	37	
13	8	Stem: Ones
14	36	Leaf: Tenths
15	0035	
16		
17		
18	9	

Stem-and-leaf display for Exercise 3.17

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.

3.19 The stem-and-leaf display of Exercise 3.18 uses only four stems. Construct a stem-and-leaf display for these data using repeated stems 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).

3.20 The accompanying observations are lengths (in yards) for a sample of golf courses recently listed by *Golf Magazine* as being among the most challenging in the United States. Construct a stem-and-leaf display, and explain why your choice of stems seems preferable to any of the other possible choices. The lengths are

6526	6770	6936	6770	6583	6464	7005	6927
6790	7209	7040	6850	6700	6614	7022	6506
6527	6470	6900	6605	6873	6798	6745	7280
7131	6435	6694	6433	6870	7169	7011	7168
6713	7051	6904	7105	7165	7050	7113	6890

3.21 Many states face a shortage of fully credentialed teachers. The percentages of teachers who are fully credentialed for each county in California were published in the *San Luis Obispo Tribune* (July 29, 2001) and are given in the following table:

County	Percentage Credentialed	County	Percentage Credentialed
Alameda	85.1	Butte	98.2
Alpine	100.0	Calaveras	97.3
Amador	97.3	Colusa	92.8
Contra Costa	87.7	Sacramento	95.3
Del Norte	98.8	San Benito	83.5
El Dorado	96.7	San Bernardino	83.1
Fresno	91.2	San Diego	96.6
Glenn	95.0	San Francisco	94.4
Humbolt	98.6	San Joaquin	86.3
Imperial	79.8	San Luis Obispo	98.1
Inyo	94.4	San Mateo	88.8
Kern	85.1	Santa Barbara	95.5
Kings	85.0	Santa Clara	84.6
Lake	95.0	Santa Cruz	89.6
Lassen	89.6	Shasta	97.5
Los Angeles	74.7	Sierra	88.5
Madera	91.0	Siskiyou	97.7
Marin	96.8	Solano	87.3
Mariposa	95.5	Sonoma	96.5
Mendicino	97.2	Stanislaus	94.4
Merced	87.8	Sutter	89.3
Modoc	94.6	Tehama	97.3
Mono	95.9	Trinity	97.5
Monterey	84.6	Tulare	87.6
Napa	90.8	Tuolomne	98.5
Nevada	93.9	Ventura	92.1
Orange	91.3	Yolo	94.6
Placer	97.8	Yuba	91.6
Plumas	95.0		
Riverside	84.5		

a. Construct a stem-and-leaf display for this data set using stems 7, 8, 9, and 10. Truncate the leaves to a single digit. Comment on the interesting features of the display.

b. Construct a stem-and-leaf display using repeated stems. Are there characteristics of the data set that are easier to see in the plot with repeated stems, or is the general shape of the two displays similar?

3.22 An article on peanut butter in *Consumer Reports* (September 1990) reported the following scores (quality ratings on a scale of 0 to 100) for various brands:

Creamy:	56	44	62	36	39	53	50	65	45	40
	56	68	41	30	40	50	56	30	22	
Crunchy:	62	53	75	42	47	40	34	62	52	50
	34	42	36	75	80	47	56	62		

Construct a comparative stem-and-leaf display, and discuss similarities and differences for the two types.

3.23 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
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.

3.24 High school dropout rates (percentages) for the period 1997–1999 for the 50 states were given in *The Chronicle of Higher Education* (August 31, 2001) and are shown in the following table:

State	Rate	State	Rate
Alabama	10	Montana	8
Alaska	7	Nebraska	8
Arizona	17	Nevada	17
Arkansas	12	New Hampshire	7
California	9	New Jersey	6
Colorado	13	New Mexico	13
Connecticut	9	New York	9
Delaware	11	North Carolina	11
Florida	12	North Dakota	5
Georgia	13	Ohio	8
Hawaii	5	Oklahoma	9
Idaho	10	Oregon	13
Illinois	9	Pennsylvania	7
Indiana	6	Rhode Island	11
Iowa	7	South Carolina	9
Kansas	7	South Dakota	8
Kentucky	11	Tennessee	12
Louisiana	11	Texas	12
Maine	7	Utah	9
Maryland	7	Vermont	6
Massachusetts	6	Virginia	8
Michigan	9	Washington	8
Minnesota	6	West Virginia	8
Mississippi	10	Wisconsin	5
Missouri	9	Wyoming	9

Note that dropout rates range from a low of 5% to a high of 17%. 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 1 could be repeated as

1. with leaves 0 and 1
1t with leaves 2 and 3
1f with leaves 4 and 5
1s with leaves 6 and 7

If there had been any dropout rates as large as 18 or 19 in the data set, we would also need to include a stem 1* to accommodate the leaves of 8 and 9.

Construct a stem-and-leaf display for this data set that uses stems 0f, 0s, 0*, 1., 1t, 1f, and 1s. 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 summary technique; it is unwieldy when the data set contains a great many observations. Frequency distributions and histograms are displays that are useful for summarizing even a large data set in a compact fashion.

▪ 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 Raccoons!

The authors of the article "Behavioral Aspects of the Raccoon Mating System: Determinants of Consortship Success" (*Animal Behaviour* [1999]: 593–601) monitored raccoons in southern Texas during three mating seasons in an effort to describe mating behavior. Twenty-nine female raccoons were observed, and the number of male partners during the time the female was accepting partners (generally 1 to 4 days each year) was recorded for each female. The resulting data were as follows:

1	3	2	1	1	4	2	4	1	1	1	3	1	1	1
1	2	2	1	1	4	1	1	2	1	1	1	1	3	

The corresponding frequency distribution is given in Table 3.1. From the frequency distribution, we can see that 18 of the female raccoons had a single partner. The corresponding relative frequency, .621, tells us that the proportion of female raccoons in the sample with a single partner was .621, or, equivalently, 62.1% of the females had a single partner. Adding the relative frequencies for the values of 1 and 2 gives

$$.621 + .172 = .793$$

indicating that 79.3% of the raccoons had 2 or fewer partners.

Table 3.1 ▪ Frequency Distribution for Number of Partners

Number of Partners	Frequency	Relative Frequency
1	18	.621
2	5	.172
3	3	.103
4	3	.103
	29	.999

Differs from 1 due to rounding

A histogram for discrete numerical data is a graph of the 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 frequencies or relative frequencies.
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, etc.). 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

Central or typical value, extent of spread or variation, general shape, location and number of peaks, and presence of gaps and outliers.

▪ Example 3.13 Revisiting Promiscuous Raccoons

The raccoon data of Example 3.12 were summarized in a frequency distribution. The corresponding histogram is shown in Figure 3.16. 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.

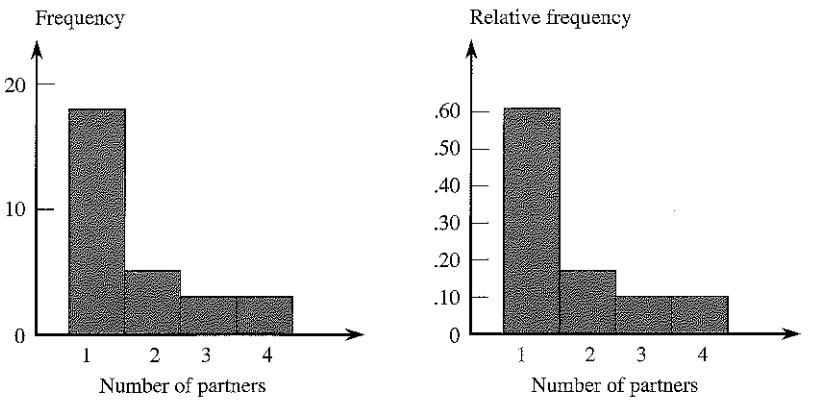


Figure 3.16 Histogram and relative frequency histogram of raccoon data.

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 Alcohol Use by College Students

Each student in a sample of 176 students at a large public university was asked about alcohol use, and the resulting data appeared in the article "Alcohol, Tobacco, and Marijuana Use: Relationships to Undergraduate Students' Creative Achievement" (*Journal of College Student Development* [1998]: 472–479). The frequency distribution in Table 3.2 summarizes the data on number of drinks consumed per week. The authors of this article chose to group the observed values (rather than list them all: 0, 1, 2, . . .), and they also created an open-ended group of "16 or more." This results in a much more compact table that still communicates one of the important features of the data: the large numbers of individuals at both the low and the high ends. Also note that we could not draw a histogram based on this frequency distribution because of the open-ended group ("16 or more"). Furthermore, because the number of possible values differs from group to group (2 in the "0 to 1" group, 4 in the "2 to 5" group, and 6 in the "10 to 15" group), even without the open-ended group, drawing a correct histogram is a bit more complicated.

Table 3.2 ■ Frequency Distribution for Number of Drinks per Week

Drinks per Week	Frequency
0 to 1	52
2 to 5	38
6 to 9	17
10 to 15	35
16 or more	34

You cannot just use frequency or relative frequency for the vertical scale. One method of drawing a histogram that would work, as long as there are no open-ended groups, is described later in this section.

▪ 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 fuel efficiency (in miles per gallon), is that there are no natural categories. The way out of this dilemma is to define our own categories. For fuel efficiency data, suppose that we mark some intervals on a horizontal miles-per-gallon measurement axis, as pictured in Figure 3.17. Each data value should fall in exactly one of these intervals. If the smallest observation was 25.3 and the largest was 29.8, we might use intervals of width 0.5, with the first interval starting at 25.0 and the last interval ending at 30.0. The resulting intervals are called **class intervals**, or just **classes**. The class intervals play the same role that the categories or individual values played in frequency distributions for categorical or discrete numerical data.

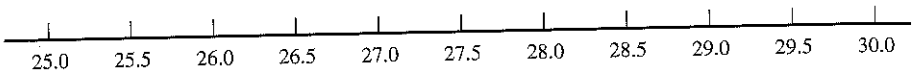


Figure 3.17 Suitable class intervals for miles-per-gallon data.

There is one further difficulty. Where should we place an observation such as 27.0, 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 our frequency distribution, a typical class will be 26.5 to <27.0, where the symbol < is a substitute for the phrase *less than*. This class will contain all observations that are greater than or equal to 26.5 and less than 27.0. The observation 27.0 would then fall in the class 27.0 to <27.5.

▪ Example 3.15 Enrollments at Public Universities

States differ widely with respect to the percentage of college students who are enrolled in public institutions. The U.S. Department of Education provided the accompanying data on this percentage for the 50 U.S. states for fall 1999. Observations for a few of the states have been identified by name.

Percentage of College Students Enrolled in Public Institutions

95 (Alaska)	81	85	80	72	73	74	79
95 (Nevada)	84	89	63	91	86	89	92
87	90	83	84	89	96 (Wyoming)	87	85
76	84	75	81	73	82	81	77	75
70	55 (New York)	56 (Pennsylvania)			87	88	82
81	84	76	80	56 (Vermont)	55 (New Hampshire)		
43 (Massachusetts)		52	62	80	82		

The smallest observation is 43 (Massachusetts) and the largest is 96 (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 3.3 displays the resulting frequency distribution, along with the relative frequencies.

Table 3.3 ▪ Frequency Distribution for Percentage of College Students Enrolled in Public Institutions

Class Interval	Frequency	Relative Frequency
40 to <50	1	.02
50 to <60	5	.10
60 to <70	2	.04
70 to <80	11	.22
80 to <90	25	.50
90 to <100	6	.12
	50	1.00

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

$$\begin{pmatrix}\text{proportion of states}\\ \text{with percentage in public}\\ \text{institutions less than 60}\end{pmatrix} = \begin{pmatrix}\text{proportion in}\\ \text{40 to <50 class}\end{pmatrix} + \begin{pmatrix}\text{proportion in}\\ \text{50 to <60 class}\end{pmatrix}$$
$$= .02 + .10 = .12\,(12\%)$$

and

$$\begin{pmatrix}\text{proportion of states}\\ \text{with percentage in public}\\ \text{institutions between 60 and 90}\end{pmatrix} = \begin{pmatrix}\text{proportion in}\\ \text{60 to <70 class}\end{pmatrix} + \begin{pmatrix}\text{proportion in}\\ \text{70 to <80 class}\end{pmatrix} + \begin{pmatrix}\text{proportion in}\\ \text{80 to <90 class}\end{pmatrix}$$
$$= .04 + .22 + .50 = .76\,(76\%)$$

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, whereas 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 summaries of the data.

▪ 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 classes and relative frequencies:

Class	0 to <25	25 to <50	50 to <75	75 to <100	100 to <125	...
Rel. freq.	.05	.10	.18	.25	.20	...

Then

$$\begin{aligned}\text{proportion of observations less than } 75 &= \text{proportion in one of the first three classes}\\ &= .05 + .10 + .18\\ &= .33\end{aligned}$$

Similarly,

$$\begin{aligned}\text{proportion of observations less than } 100 &= .05 + .10 + .18 + .25\\ &= .33 + .25\\ &= .58\end{aligned}$$

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. These calculations can also be done for discrete data (e.g., the proportion of observations that are at most 5, at most 6, etc.). The use of cumulative relative frequencies is illustrated in Example 3.16.

▪ Example 3.16 Strength of Aircraft Welds

The strength of welds used in aircraft construction has been of great concern to aeronautical engineers in recent years. Table 3.4 gives a frequency distribution for shear strengths (force in pounds required to break the weld) of ultrasonic spot welds ("Comparison of Properties of Joints Prepared by Ultrasonic Welding and Other Means," *Journal of Aircraft* [1983]: 552–556).

The proportion of welds with strength values less than 5400 is .85 (i.e., 85% of the observations are below 5400). What about the proportion of observations less than 4700? Because 4700 is not a class boundary, we must make an educated guess. The value 4700 is halfway between the boundaries of the 4600–4800 class; let's estimate that half of the relative frequency of .14 for this class, or .07, belongs in the 4600–4700 range. Thus,

$$\text{estimate of proportion less than } 4700 = .01 + .02 + .09 + .07 = .19$$

Table 3.4 ▪ Frequency Distribution with Cumulative Relative Frequencies

Class Interval	Frequency	Relative Frequency	Cumulative Relative Frequency
4000 to <4200	1	.01	.01
4200 to <4400	2	.02	.03
4400 to <4600	9	.09	.12
4600 to <4800	14	.14	.26
4800 to <5000	17	.17	.43
5000 to <5200	22	.22	.65
5200 to <5400	20	.20	.85
5400 to <5600	7	.07	.92
5600 to <5800	7	.07	.99
5800 to <6000	1	.01	1.00
	100	1.00	

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

estimate of proportion less than 4700 = .12 + .07 = .19

Similarly, because 5250 is one-fourth of the way from 5200 to 5400,

estimate of proportion less than 5250 = .65 + .25(.20) = .70

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 weld strength data of Example 3.16, the plotted points would be

(4200, .01) (4400, .03) (4600, .12) (4800, .26) (5000, .43)
(5200, .65) (5400, .85) (5600, .92) (5800, .99) (6000, 1.00)

One additional point, the pair (lower endpoint of first interval, 0), is also included in the plot (for the weld strength data, this would be the point (4000, 0)), and then points are connected by line segments. Figure 3.18 shows the cumulative relative frequency plot for the weld strength 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?

For example, to determine the approximate proportion of the weld strength values that are smaller than 5700, we would follow a vertical line up from 5700 on

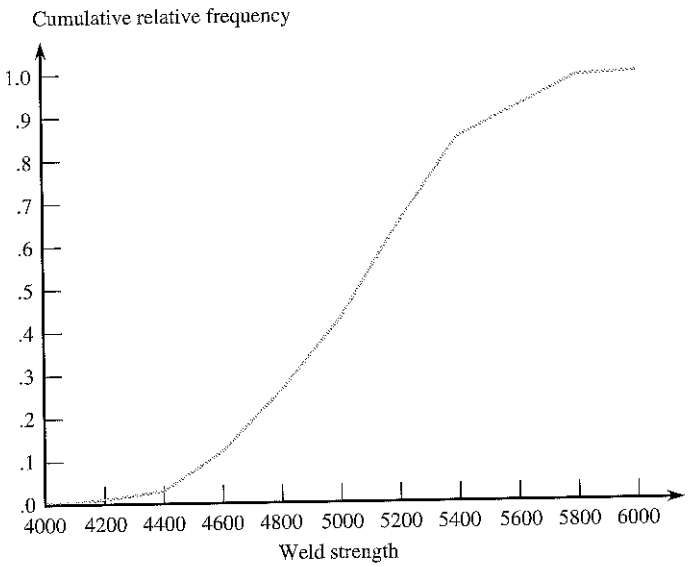


Figure 3.18 Cumulative relative frequency plot for the weld strength data of Example 3.16.

the x axis and then read across to obtain the corresponding cumulative relative frequency, as illustrated in Figure 3.19(a). Approximately .94, or 94%, of the weld strengths are smaller than 5700. Similarly, to find the weld strength value that separates the smallest 30% of the weld strengths from the larger values, start at .30 on the cumulative relative frequency axis and move across and then down to find the corresponding weld strength value, as shown in Figure 3.19(b). Approximately 30% of the weld strengths are smaller than 4850.

▪ Histograms for Continuous Numerical Data

In Example 3.15, the class intervals in the frequency distribution were all of equal width. When this is the case, 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 boundaries). The height of each rectangle is the frequency or relative frequency of the corresponding class interval.

What to Look For
Central or typical value, extent of spread or variation, general shape, location and number of peaks, and presence of gaps and outliers.

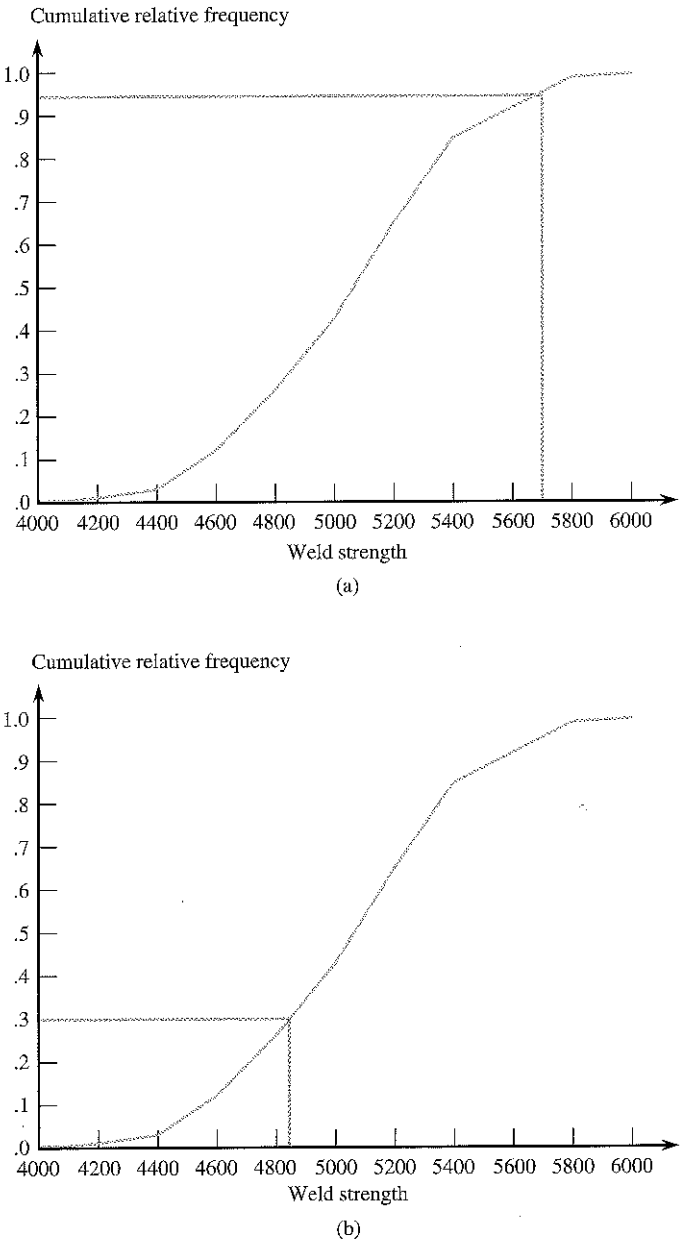


Figure 3.19 Approximating weld strength using the cumulative relative frequency plot. (a) Determining the approximate proportion of weld strength values less than 5700; (b) finding the weld strength value that separates the smallest 30% of weld strengths from the larger values.

▪ Example 3.17 Mercury Contamination

StatisticsNow™

Explore this example with technology. Download your technology manual at the StatisticsNow website or from your CD.

Mercury contamination is a serious environmental concern. Mercury levels are particularly high in certain types of fish. Citizens of the Republic of Seychelles, a group of islands in the Indian Ocean, are among those who consume the most fish in the world. The article "Mercury Content of Commercially Important Fish of the Seychelles, and Hair Mercury Levels of a Selected Part of the Population" (*Environmental Research* [1983]: 305–312) reported the following observations on mercury content (in parts per million) in the hair of 40 fishermen:

13.26	32.43	18.10	58.23	64.00	68.20	35.35	33.92	23.94	18.28
22.05	39.14	31.43	18.51	21.03	5.50	6.96	5.19	28.66	26.29
13.89	25.87	9.84	26.88	16.81	37.65	19.63	21.82	31.58	30.13
42.42	16.51	21.16	32.97	9.84	10.64	29.56	40.69	12.86	13.80

A reasonable choice for class intervals is to start the first interval at 0 and set the interval width as 10. The resulting frequency distribution is displayed in Table 3.5, and the corresponding histogram appears in Figure 3.20. If it were not for the slight dip in the interval 50 to <60, the histogram would have a single peak; this dip might well disappear with a larger sample size. The upper or right end of the histogram is much more stretched out than the lower or left end. Typical mercury content is somewhere between 20 and 30, but the data exhibit a substantial amount of variability about the center.

Table 3.5 ▪ Frequency Distribution for Hair Mercury Content (ppm) of Seychelles Fishermen

Class Interval	Frequency	Relative Frequency
0 to <10	5	.125
10 to <20	11	.275
20 to <30	10	.250
30 to <40	9	.225
40 to <50	2	.050
50 to <60	1	.025
60 to <70	2	.050
	40	1.000

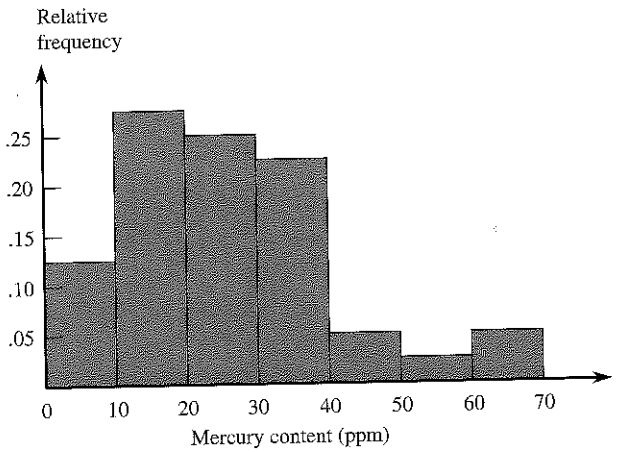


Figure 3.20 Histogram for hair mercury content of Seychelles fishermen.

Figure 3.21 shows a histogram generated by the MINITAB statistical package. Notice that the classes are different from the ones we chose and that the centers of the intervals rather than the class boundaries are marked. There is now a gap toward the upper end of the data; this discrepancy in the two graphical displays may well be attributable to the small sample size. With larger samples, the choice of intervals has less impact on the appearance of the histogram.

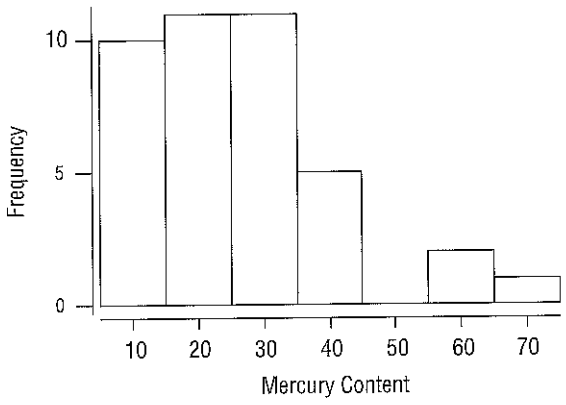


Figure 3.21 Histogram for hair mercury content of Seychelles fishermen created using MINITAB.

▪ **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 set, with only a few outlying, 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 (0 frequency). 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. Neither choice yields an informative description of the distribution. In such a situation, it is best to use a combination of relatively wide class intervals where there are few data points and relatively shorter intervals where there are many data points.

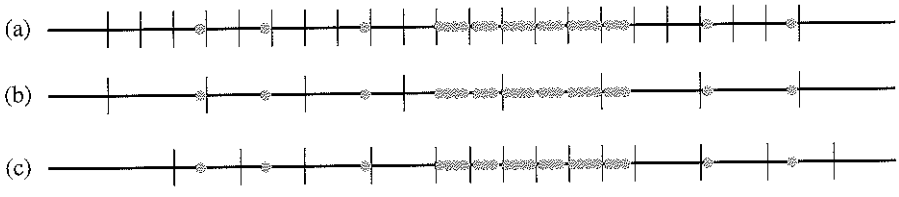


Figure 3.22 Three choices of class intervals for a data set with outliers: (a) many short intervals of equal width; (b) a few wide intervals of equal width; (c) intervals of unequal width.

▪ **Constructing a Histogram for Continuous Data When Class Widths Are Unequal**

When class widths are unequal, frequencies or relative frequencies should *not* be used on the vertical axis of a histogram. Instead, the height of each rectangle is determined by the **density**. The density is given by

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

The vertical axis is called the **density scale**; it should be marked so that each rectangle can be drawn to the calculated height.

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 identical; the denominator is then the same for each density calculation. The resulting histogram will look exactly like the one based on relative frequencies, except for the vertical scaling. When the intervals are of equal width, the extra arithmetic required to obtain the densities is unnecessary.

▪ Example 3.18 Misreporting Grade Point Averages

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 gross errors. Because of this, a frequency distribution based on unequal class widths gives an informative yet concise summary. Table 3.6 displays such a distribution based on classes with boundaries at −2.0, −0.4, −0.2, −0.1, 0, 0.1, 0.2, 0.4, and 2.0.

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	0.2	0.275
−0.2 to <−0.1	.097	0.1	0.970
−0.1 to <0	.210	0.1	2.100
0 to <0.1	.189	0.1	1.890
0.1 to <0.2	.139	0.1	1.390
0.2 to <0.4	.116	0.2	0.580
0.4 to <2.0	.171	1.6	0.107

Figure 3.23 displays two histograms based on this frequency distribution. The histogram in Figure 3.23(a) is correctly drawn, with height equal to relative frequency divided by interval width. 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.

The formula for rectangle height given previously implies that

$$\text{relative frequency} = (\text{rectangle height})(\text{class width}) = \text{area of rectangle}$$

That is, a histogram can always be drawn so that the area of each rectangle is the relative frequency of the corresponding class (thus the total area of all rectangles is 1).

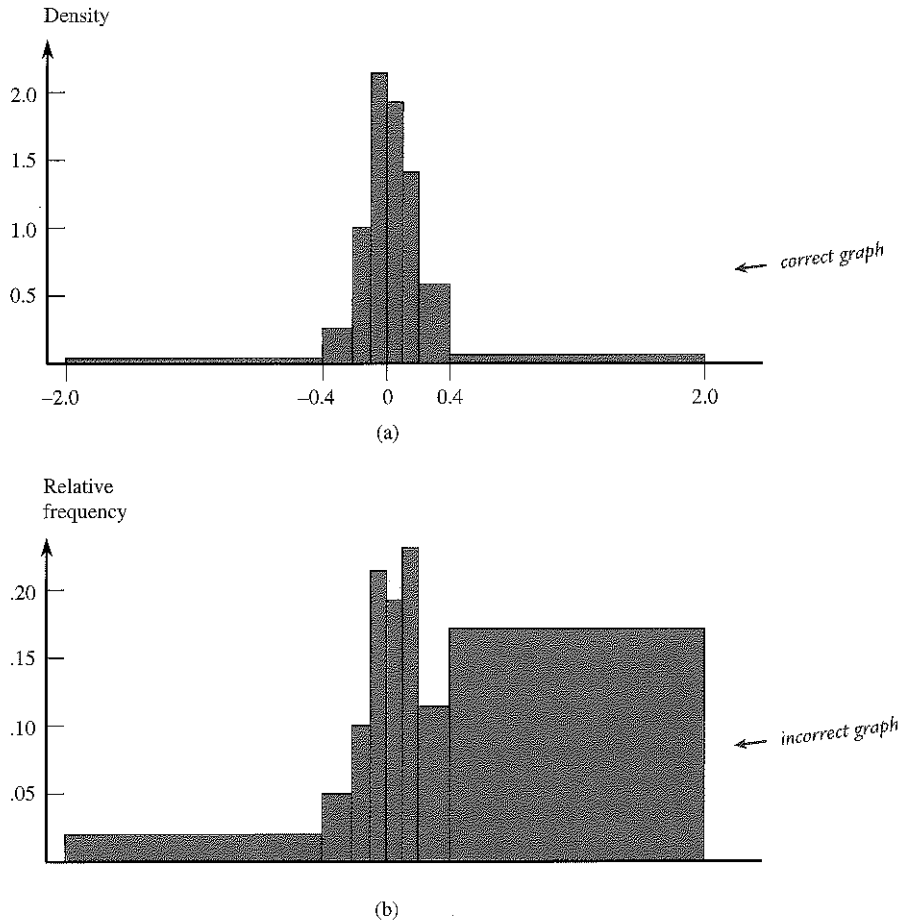


Figure 3.23 Histograms for errors in reporting GPA: (a) a correct picture (height = density); (b) an incorrect picture (height = relative frequency).

▪ Histogram Shapes

The general shape of a histogram is an important characteristic. 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.

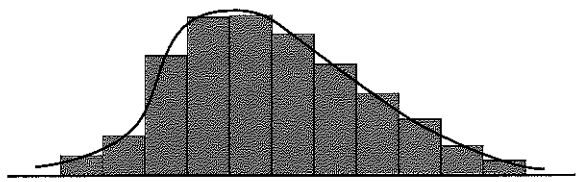


Figure 3.24 Approximating a histogram with a smooth curve.

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

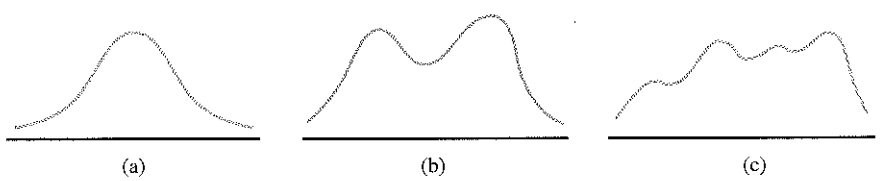


Figure 3.25 Smoothed histograms with various numbers of modes: (a) unimodal; (b) bimodal; (c) multimodal.

Bimodality can occur 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 hr) and another for those cars traveling up the coast highway (3.5–4 hr). 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 rather rare.

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

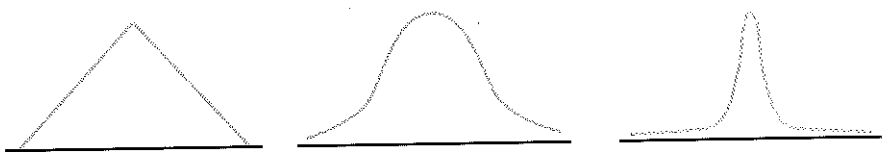


Figure 3.26 Several symmetric unimodal smoothed histograms.

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**. 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**. If, on the other hand, the lower tail is much longer than the upper tail, the histogram is **negatively 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.

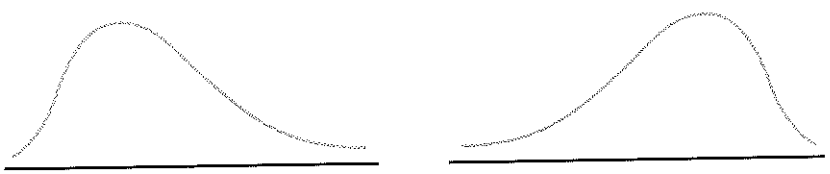


Figure 3.27 Two examples of skewed smoothed histograms: (a) positive skew; (b) negative skew.

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 (e.g., characteristics such as blood pressure, brain weight, adult male height, adult female height, and IQ score). Here we briefly mention several of the most important qualitative properties of such a curve, postponing a more detailed discussion until Chapter 7. A normal curve is not only symmetric but also bell

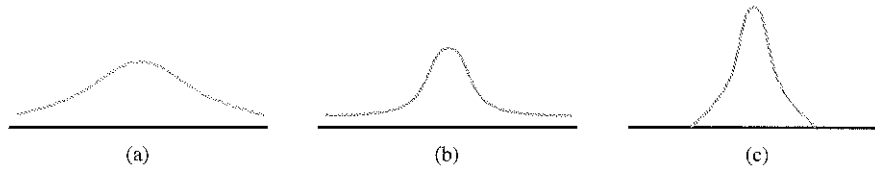


Figure 3.28 Three examples of bell-shaped histograms: (a) normal; (b) heavy-tailed; (c) light-tailed.

shaped; it looks like the curve in Figure 3.28(a). However, not all bell-shaped curves are normal. From the top of the bell, the height of the curve decreases at a well-defined rate when moving into either tail. (This rate of decrease is specified by a certain mathematical function.)

A curve with tails that do not decline as rapidly as the tails of a normal curve is said to specify a **heavy-tailed** distribution (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 when the population distribution is approximately normal (i.e., they result in accurate conclusions) do poorly when the population distribution is heavy tailed.

▪ Do Sample Histograms Resemble the Population Histogram?

Sample data are usually collected to make inferences about a population. The resulting conclusions may be in error if the sample is somehow 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 these 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 3.19 illustrates such variability in histogram shapes.

▪ Example 3.19 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

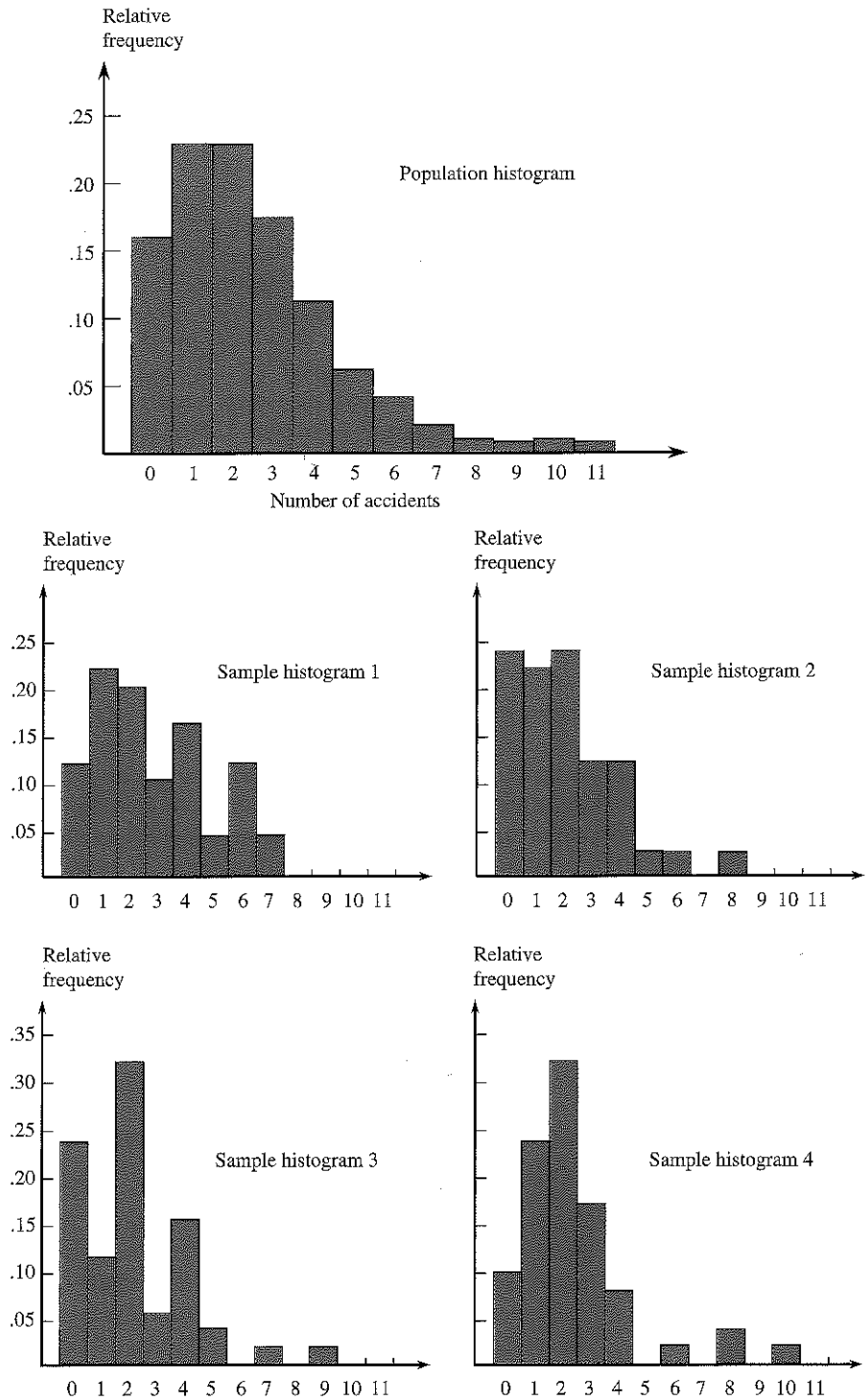


Figure 3.29
Comparison of population and sample histograms for number of accidents.

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

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 from this population. The five histograms 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 incorporated into conclusions based on inferential methods.

▪ Exercises 3.25–3.40

Mentor Access one-on-one tutoring from a statistics expert at **http://1pass.thomson.com**.

3.25 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?
e. Do you notice anything especially interesting about the frequencies and relative frequencies?
Note: The investigators proposed statistical models from which the number of impairments could be predicted from various patient characteristics and symptoms.

3.26 The trace element zinc is an important dietary constituent, partly because it aids in the maintenance of the immune system. The accompanying data on zinc intake (in milligrams) for a sample of 40 patients with rheumatoid arthritis were read from a graph in the article "Plasma Zinc and Copper Concentrations in Rheumatoid Arthritis: Influence of Dietary Factors and Disease Activity" (*American Journal of Clinical Nutrition* [1991]: 1082–1086):

8.0	12.9	13.0	8.9	10.1	17.3	11.1	10.9
6.2	8.1	8.8	10.4	15.7	13.6	19.3	9.9
8.5	11.1	10.7	8.8	10.7	6.8	7.4	4.8
11.8	13.0	9.5	8.1	6.9	11.5	11.2	13.6
4.9	18.8	15.7	10.8	10.7	11.5	16.1	9.9

a. Use class intervals of 3 to <6, 6 to <9, 9 to <12, 12 to <15, 15 to <18, and 18 to <21 to construct a table that includes the class intervals, frequencies, and relative frequencies.
b. Use the table constructed in Part (a) to determine the proportion of individuals whose intake is less than 12 and the proportion whose intake is between 6 and 15.
c. Construct a histogram for this data set, and comment on the key features of the histogram.

3.27 *USA Today* (July 2, 2001) gave the following information regarding cell phone use for men and women:

Average Number of Minutes Used per Month	Relative Frequency	
	Men	Women
0 to <200	.56	.61
200 to <400	.18	.18
400 to <600	.10	.13
600 to <800	.16	.08

a. Construct a relative frequency histogram for average number of minutes used per month for men. How would you describe the shape of this histogram?
b. Construct a relative frequency histogram for average number of minutes used per month for women. Is the distribution for average number of minutes used per month similar for men and women? Explain.
c. What proportion of men average less than 400 minutes per month?
d. Estimate the proportion of men that average less than 500 minutes per month.
e. Estimate the proportion of women that average 450 minutes or more per month.

3.28 The article "Associations Between Violent and Nonviolent Criminality" (*Multivariate Behavioral Research* [1981]: 237–242) reported the number of previous convictions for 283 adult males arrested for felony offenses. The following frequency distribution is a summary of the data given in the paper:

Number of Previous Convictions	Frequency
0	0
1	16
2	27
3	37
4	46
5	36
6	40
7	31
8	27
9	13
10	8
11	2

Draw the histogram corresponding to this frequency distribution, and comment on its shape.

3.29 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.
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 min?
ii. Approximately what proportion of commute times were greater than 22 min?
iii. What is the approximate commute time value that separates the shortest 50% of commute times from the longest 50%?

3.30 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	

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?

3.31 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 it would be inappropriate to construct a histogram using relative frequencies to determine the height of the bars in the histogram?
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.

3.32 The behavior of children watching television has been a much-studied phenomenon. The same attention has been given to children engaged in playing with toys. The paper "A Temporal Analysis of Free Toy Play and Distractibility in Young Children" (*Journal of Experimental Child Psychology* [1991]: 41–69) reported the following data on play-episode lengths (in seconds) for a particular 5-year-old boy:

Class	Frequency
0 to <5	54
5 to <10	44
10 to <15	28
15 to <20	21
20 to <40	31
40 to <60	15
60 to <90	16
90 to <120	5
120 to <180	8

a. Display this information in a histogram.
b. What proportion of episodes lasted at least 20 sec?
c. Roughly what proportion of episodes lasted between 40 and 75 sec?

3.33 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.

3.34 The results of the 1990 census included a state-by-state listing of population density. The following table gives the number of people per square mile for each of the 50 states:

State	Number of People per Square Mile
Alabama	79.6
Alaska	1.0
Arizona	32.3
Arkansas	45.1
California	190.8
Colorado	31.8
Connecticut	678.4
Delaware	340.8
Florida	239.6
Georgia	111.9
Hawaii	172.5
Idaho	12.2
Illinois	205.6
Indiana	154.6
Iowa	49.7
Kansas	30.3
Kentucky	92.8
Louisiana	96.9
Maine	39.8
Maryland	489.2
Massachusetts	767.6
Michigan	163.6
Minnesota	55.0
Mississippi	54.9
Missouri	74.3
Montana	5.5
Nebraska	20.5
Nevada	10.9
New Hampshire	123.7
New Jersey	1042.0
New Mexico	12.5
New York	381.0
North Carolina	136.1
North Dakota	9.3
Ohio	264.9
Oklahoma	45.8
Oregon	29.6
Pennsylvania	265.1
Rhode Island	960.3
South Carolina	115.8
South Dakota	9.2
Tennessee	118.3
Texas	64.9
Utah	21.0
Vermont	60.8
Virginia	156.3
Washington	73.1
West Virginia	74.5
Wisconsin	90.1
Wyoming	4.7

a. Construct a relative frequency distribution for state population density.
b. In your relative frequency distribution, did you use class intervals of equal widths? Why or why not?
c. Use the relative frequency distribution to give an approximate value for the proportion of states that have a population density of more than 100 people per square mile. Is this value close to the actual value?

3.35 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	32							

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?

3.36 The clearness index was determined for the skies over Baghdad for each of the 365 days during a particular year ("Contribution to the Study of the Solar Radiation Climate of the Baghdad Environment," *Solar Energy* [1990]: 7–12). The accompanying table summarizes the resulting data:

Clearness Index	Number of Days (frequency)
0.15 to <0.25	8
0.25 to <0.35	14
0.35 to <0.45	28
0.45 to <0.50	24
0.50 to <0.55	39
0.55 to <0.60	51
0.60 to <0.65	106
0.65 to <0.70	84
0.70 to <0.75	11

a. Determine the relative frequencies and draw the corresponding histogram. (Be careful here — the intervals do not all have the same width.)

b. Cloudy days are those with a clearness index smaller than 0.35. What proportion of the days were cloudy?

c. Clear days are those for which the index is at least 0.65. What proportion of the days were clear?

3.37 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 to run the first 5 km and the time 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 accompanying histogram 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?

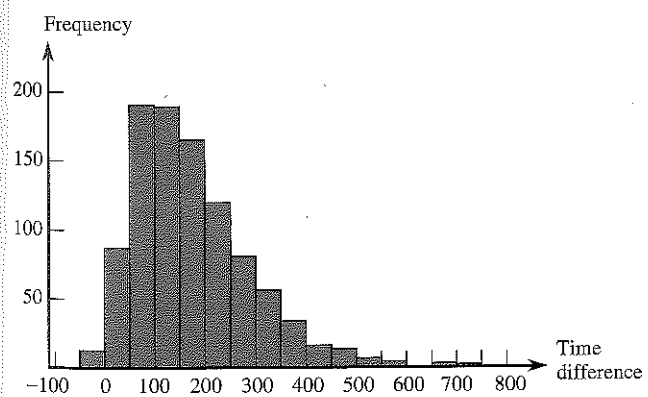


Figure for Exercise 3.37

3.38 Disparities among welfare payments by different states have been the source of much political controversy. The accompanying table reports average payment per person (in dollars) in the Aid to Families with Dependent Children Program for the 1990 fiscal year. Construct a relative frequency distribution for these data using equal interval widths. Draw the histogram corresponding to your frequency distribution.

State	Average Welfare Payment ($)
Alaska	244.90
California	218.31
Arizona	93.57
Montana	114.95
Texas	56.79
Nebraska	115.15
Minnesota	171.75
Arkansas	65.96
Alabama	39.62
Illinois	112.28
Indiana	92.43
New Hampshire	164.20
Rhode Island	179.37
New Jersey	121.99
Delaware	113.66
North Carolina	91.95
Florida	95.43
Washington	160.41
Idaho	97.93
Utah	118.36
Colorado	111.20
Oklahoma	96.98
South Dakota	95.52
Iowa	129.58
Louisiana	55.81
Tennessee	65.93

(*continued*)