

CHAPTER

1

Exploring Data

1.1 Displaying Distributions with Graphs

1.2 Describing Distributions with Numbers

Chapter Review

CASE STUDY

Nielsen ratings

What does it mean to say that a TV show was ranked number 1? The Nielsen Media Research company randomly samples about 5100 households and 13,000 individuals each week. The TV viewing habits of this sample are captured by metering equipment, and data are sent automatically in the middle of the night to Nielsen. Broadcasters and companies that want to air commercials on TV use the data on who is watching TV and what they are watching. The results of this data gathering appear as ratings on a weekly basis. For more information on the Nielsen TV ratings, go to www.nielsenmedia.com, and click on "Inside TV Ratings." Then under "Related," select "What Are TV Ratings?"

Here are the top prime-time shows for viewers aged 18 to 49 during the week of November 22–28, 2004.



Show	Network	Viewers (millions)
1. <i>Desperate Housewives</i>	ABC	16.2
2. <i>CSI</i>	CBS	10.9
3. <i>CSI: Miami</i>	CBS	10.5
4. <i>Extreme Makeover: Home Edition</i>	ABC	9.7
5. <i>Two and a Half Men</i>	CBS	8.8
6. <i>Without a Trace</i>	CBS	8.2
7. <i>Raymond</i>	CBS	8.0
8. <i>Law & Order: SVU</i>	NBC	7.8
9. <i>Monday Night Football</i>	ABC	7.8
10. <i>Survivor: Vanuatu</i>	CBS	7.8
11. <i>Seinfeld Story</i>	NBC	7.6
12. <i>Boston Legal</i>	ABC	7.4
13. <i>Apprentice</i>	NBC	7.1
14. <i>Fear Factor</i>	NBC	6.5
15. <i>Amazing Race</i>	CBS	6.1
16. <i>CSI: NY</i>	CBS	6.1
17. <i>NFL Monday Showcase</i>	ABC	5.7
18. <i>According to Jim</i>	ABC	5.5
19. <i>60 Minutes</i>	CBS	5.4
20. <i>Biggest Loser</i>	NBC	5.4

Source: *USA Today*, December 2, 2004.

Which network is winning the ratings battle? At the end of this chapter, you will be asked to use what you have learned to answer this question.



Activity 1A

How fast is your heart beating?

Materials: Clock or watch with second hand

A person's pulse rate provides information about the health of his or her heart. Would you expect to find a difference between male and female pulse rates? In this activity, you and your classmates will collect some data to try to answer this question.

1. To determine your pulse rate, hold the fingers of one hand on the artery in your neck or on the inside of your wrist. (The thumb should not be used, because there is a pulse in the thumb.) Count the number of pulse beats in one minute. As Jeremy notes in the cartoon above, you need sufficient data; so do this three times, and calculate your *average* individual pulse rate. Why is doing this three times better than doing it once?
2. Record the pulse rates for the class in a table, with one column for males and a second column for females. Are there any unusual pulse rates?
3. For now, simply calculate the average pulse rate for the males and the average pulse rate for the females, and compare.

Introduction

When you go to the movie theater, you see previews of upcoming movies. Similarly, our goal in the Preliminary Chapter was to give you a taste of what statistics is about and a sense of what lies ahead in this book. Now we are ready for the details. In this chapter, we will explore ways to describe data, both graphically and numerically. Initially, we will focus on ways to plot our data. We will learn how to construct histograms, stemplots, and boxplots, and how to decide which plot is best for different sets of circumstances. We'll talk about distributions, because

distributions will come up again and again during our study of statistics. We'll also discuss ways to describe distributions numerically, including measures of center and spread. Our recurring theme throughout will be looking for patterns and departures from patterns.

1.1 Displaying Distributions with Graphs

Graphs for Categorical Variables

Recall our distinction between quantitative and categorical variables in the Preliminary Chapter. The values of a categorical variable are labels for the categories, such as "female" and "male." The distribution of a categorical variable lists the categories and gives either the count or the percent of individuals who fall in each category. The next two examples show you how to use graphical displays to examine the distribution of a categorical variable.

Example 1.1 *Radio station formats* Bar graphs and pie charts

The radio audience rating service Arbitron places the country's 13,838 radio stations into categories that describe the kind of programs they broadcast. Here is the distribution of station formats:¹

Format	Count of stations	Percent of stations
Adult contemporary	1,556	11.2
Adult standards	1,196	8.6
Contemporary hit	569	4.1
Country	2,066	14.9
News/Talk/Information	2,179	15.7
Oldies	1,060	7.7
Religious	2,014	14.6
Rock	869	6.3
Spanish language	750	5.4
Other formats	1,579	11.4
Total	13,838	99.9

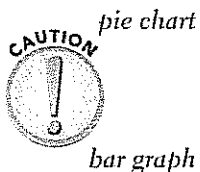
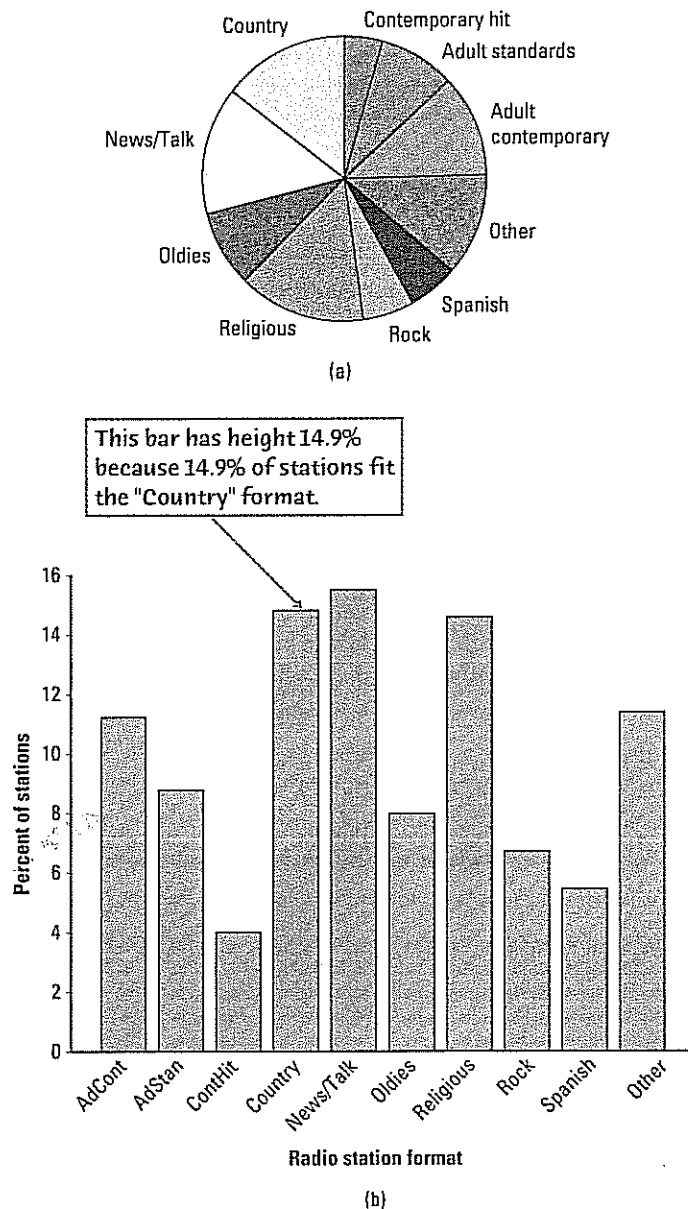
It's a good idea to check data for consistency. The counts should add to 13,838, the total number of stations. They do. The percents should add to 100%. In fact, they add to 99.9%. What happened? Each percent is rounded to the nearest tenth. The exact percents would add to 100, but the rounded percents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

roundoff error

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

Figure 1.1

(a) Pie chart of radio stations by format. (b) Bar graph of radio stations by format.



Pie charts are awkward to make by hand, but software will do the job for you. A pie chart must include all the categories that make up a whole. Use a pie chart only when you want to emphasize each category's relation to the whole.

We need the "Other formats" category in Example 1.1 to complete the whole (all radio stations) and allow us to make a pie chart. *Bar graphs* are easier to make and also easier to read, as Figure 1.1(b) illustrates. Bar graphs are more flexible than

pie charts. Both graphs can display the distribution of a categorical variable, but a bar graph can also compare any set of quantities that are measured in the same units. Bar graphs and pie charts help an audience grasp data quickly.

Example 1.2***Do you listen while you walk?***

When pie charts won't work

Portable MP3 music players, such as the Apple iPod, are popular—but not equally popular with people of all ages. Here are the percents of people in various age groups who own a portable MP3 player:²

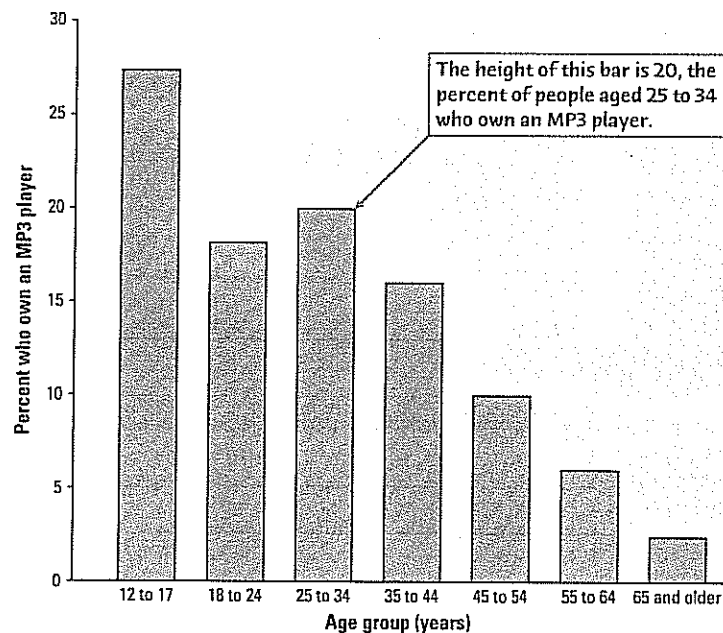
Age group (years)	Percent owning an MP3 player
12–17	27
18–24	18
25–34	20
35–44	16
45–54	10
55–64	6
65+	2

It's clear that MP3 players are popular mainly among young people.

We can't make a pie chart to display these data. Each percent in the table refers to a different age group, not to parts of a single whole. The bar graph in Figure 1.2 compares the seven age groups. We see at a glance that MP3 player ownership generally declines with age.

Figure 1.2

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



There is one question that you should always ask when you look at data, as the following example illustrates.

Example 1.3**MP3 downloads**

Do the data tell you what you want to know?



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

We now turn to the kinds of graphs that are used to describe the distribution of a quantitative variable. We will explain how to make the graphs by hand, because knowing this helps you understand what the graphs show. However, making graphs by hand is so tedious that software is almost essential for effective data analysis unless you have just a few observations.

Stemplots

A **stemplot** (also called a stem-and-leaf plot) gives a quick picture of the shape of a distribution while including the actual numerical values in the graph. Stemplots work best for small numbers of observations that are all greater than 0.

Stemplot

To make a stemplot:

1. Separate each observation into a **stem**, consisting of all but the final (right-most) digit, and a **leaf**, the final digit. Stems may have as many digits as needed, but each leaf contains only a single digit.
2. Write the stems in a vertical column with the smallest at the top, and draw a vertical line at the right of this column.
3. Write each leaf in the row to the right of its stem, in increasing order out from the stem.

Example 1.4**Literacy in Islamic countries**

Constructing and interpreting a stemplot

The Islamic world is attracting increased attention in Europe and North America. Table 1.1 shows the percent of men and women at least 15 years old who were literate in 2002 in the major Islamic nations. We omitted countries with populations of less than 3 million. Data for a few nations, such as Afghanistan and Iraq, are not available.³

Table 1.1 Literacy rates in Islamic nations

Country	Female percent	Male percent	Country	Female percent	Male percent
Algeria	60	78	Morocco	38	68
Bangladesh	31	50	Saudi Arabia	70	84
Egypt	46	68	Syria	63	89
Iran	71	85	Tajikistan	99	100
Jordan	86	96	Tunisia	63	83
Kazakhstan	99	100	Turkey	78	94
Lebanon	82	95	Uzbekistan	99	100
Libya	71	92	Yemen	29	70
Malaysia	85	92			

Source: United Nations data found at www.earthtrends.wri.org.

To make a stemplot of the percents of females who are literate, use the first digits as stems and the second digits as leaves. Algeria's 60% literacy rate, for example, appears as the leaf 0 on the stem 6. Figure 1.3 shows the steps in making the plot.

Figure 1.3

Making a stemplot of the data in Example 1.4. (a) Write the stems. (b) Go through the data and write each leaf on the proper stem. For example, the values on the 8 stem are 86, 82, and 85 in the order of Table 1.1. (c) Arrange the leaves on each stem in order out from the stem. The 8 stem now has leaves 2, 5, and 6.

2	2	9	2	9
3	3	1 8	3	1 8
4	4	6	4	6
5	5		5	
6	6	0 3 3	6	0 3 3
7	7	1 1 0 8	7	0 1 1 8
8	8	6 2 5	8	2 5 6
9	9	9 9 9	9	9 9 9
(a)	(b)	(c)		

The overall pattern of the stemplot is irregular, as is often the case when there are only a few observations. There do appear to be two *clusters* of countries. The plot suggests that we might want to investigate the variation in literacy. For example, why do the three central Asian countries (Kazakhstan, Tajikistan, and Uzbekistan) have very high literacy rates?

*back-to-back
stemplot*

When you wish to compare two related distributions, a *back-to-back stemplot* with common stems is useful. The leaves on each side are ordered out from the common stem. Here is a back-to-back stemplot comparing the distributions of female and male literacy rates in the countries of Table 1.1.

Female		Male
9	2	
8 1	3	
6	4	
	5	0
3 3 0	6	8 8
8 1 1 0	7	0 8
6 5 2	8	3 4 5 9
9 9 9	9	2 2 4 5 6
	10	0 0 0

The values on the left are the female percents, as in Figure 1.3, but ordered out from the stem, from right to left. The values on the right are the male percents. It is clear that literacy is generally higher among males than among females in these countries.

*splitting stems**trimming*

Stemplots do not work well for large data sets where each stem must hold a large number of leaves. Fortunately, there are two modifications of the basic stemplot that are helpful when plotting the distribution of a moderate number of observations. You can double the number of stems in a plot by *splitting stems* into two: one with leaves 0 to 4 and the other with leaves 5 through 9. When the observed values have many digits, *trimming* the numbers by removing the last digit or digits before making a stemplot is often best. You must use your judgment in deciding whether to split stems and whether to trim, though statistical software will often make these choices for you. Remember that the purpose of a stemplot is to display the shape of a distribution. Here is an example that makes use of both of these modifications.

Example 1.5**Virginia college tuition**
Trimming and splitting stems

Tuition and fees for the 2005–2006 school year for 37 four-year colleges and universities in Virginia are shown in Table 1.2. In addition, there are 23 two-year community colleges that each charge \$2135. The data for these 60 schools were entered into a Minitab worksheet and the following stemplot was produced.

Let's see how the software has simplified and then plotted the data. The last leaf is the cost of tuition and fees for the University of Richmond (UR). Notice that the "Leaf Unit" is given as 1000. The actual figure for UR is \$34,850. In the stemplot, the stem is 3 and the leaf is 4. Since the leaf unit is given as 1000, the plotted value for UR is then \$34,000. Minitab has truncated, or chopped off, the last three digits, leaving 34 thousand, and this truncated number is plotted. This modification is called "trimming." Notice that the number is not rounded off.


```
0 2222222222222222222229
1 0122223444444556667788999
2 1111222225
3 4
```

Exercises

Material	Weight (million tons)	Percent of total
Food scraps	25.9	11.2
Glass	12.8	5.5
Metals	18.0	7.8
Paper, paperboard	86.7	37.4
Plastics	24.7	10.7
Rubber, leather, textiles	15.8	6.8
Wood	12.7	5.5
Yard trimmings	27.7	11.9
Other	7.5	3.2
Total	231.9	100.0

- 1.2 Do the data tell you what you want to know?** To help you plan advertising for a Web site for downloading MP3 music files, you want to know what percent of owners of portable MP3 players are 18 to 24 years old. The data in Example 1.2 do *not* tell you what you want to know. Why not?

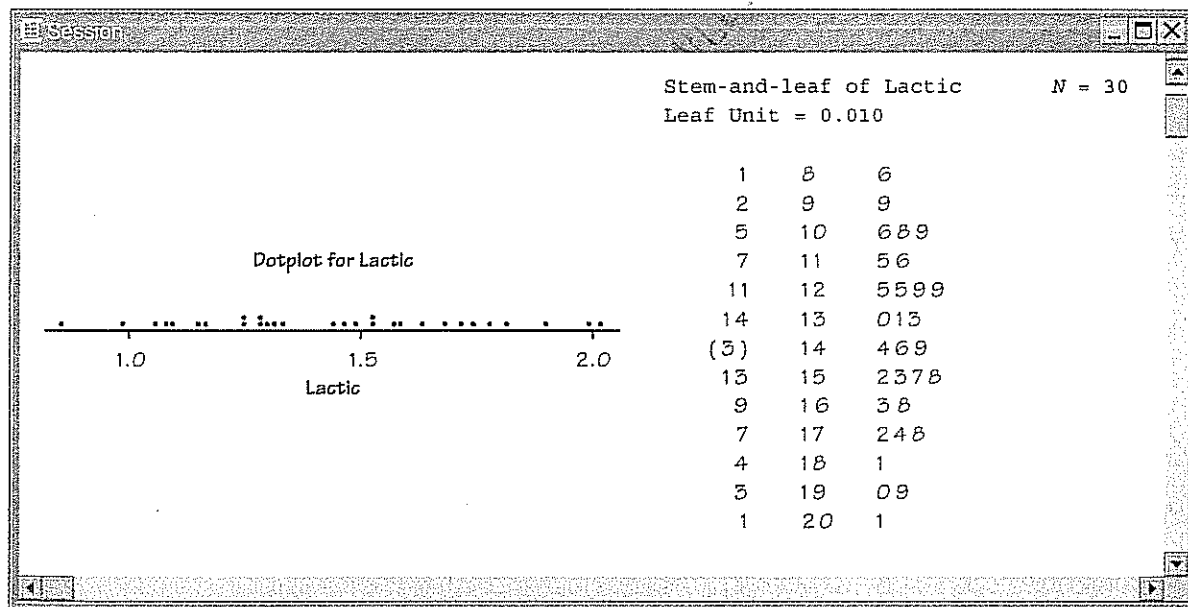
1.3 Cheese and chemistry As cheddar cheese matures, a variety of chemical processes take place. The taste of mature cheese is related to the concentration of several chemicals in the final product. In a study of cheddar cheese from the Latrobe Valley of Victoria, Australia, samples of cheese were analyzed for their chemical composition. The final concentrations of lactic acid in the 30 samples, as a multiple of their initial concentrations, are given below.⁵

0.86	1.53	1.57	1.81	0.99	1.09	1.29	1.78	1.29	1.58
1.68	1.90	1.06	1.30	1.52	1.74	1.16	1.49	1.63	1.99
1.15	1.33	1.44	2.01	1.31	1.46	1.72	1.25	1.08	1.25

A dotplot and a stemplot from the Minitab statistical software package are shown in Figure 1.5. Recall that dotplots were discussed in the Preliminary Chapter.

Figure 1.5

Minitab dotplot and stemplot for cheese data, for Exercise 1.3.



- Which plot does a better job of summarizing the data? Explain why.
- What do the numbers in the left column in the stemplot tell us? How does Minitab identify the row that contains the center of the distribution?
- The final concentration of lactic acid in one of the samples stayed approximately the same (as its initial concentration). Identify this sample in both plots.

1.4 Virginia college tuitions Tuitions and fees for the 2005–2006 school year for 60 two- and four-year colleges and universities in Virginia are shown in Table 1.2 (page 45) and the stemplot in Figure 1.4 (page 45) in Example 1.5. They ranged from a low of \$2135 for the two-year community colleges to a high of \$34,850 for the University of Richmond.

- Which stem and leaf represent Liberty University? Which stem and leaf represent Virginia State University? Identify the colleges represented in the row 2 1111.

(b) There is a string of 23 twos at the top of the stemplot. Which colleges do these numbers refer to? The stem for these entries is 0. What range of numbers would be plotted with this stem?

1.5 DRP test scores There are many ways to measure the reading ability of children. One frequently used test is the Degree of Reading Power (DRP). In a research study on third-grade students, the DRP was administered to 44 students.⁶ Their scores were

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

Display these data graphically. Write a paragraph describing the distribution of DRP scores.



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

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

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

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

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

Histograms

histogram

Stemplots display the actual values of the observations. This feature makes stemplots awkward for large data sets. Moreover, the picture presented by a stemplot divides the observations into groups (stems) determined by the number system rather than by judgment. Histograms do not have these limitations. A **histogram** breaks the range of values of a variable into *classes* and displays only the count or percent of the observations that fall into each class. You can choose any convenient number of classes, but you should *always choose classes of equal width*. Histograms are slower to construct by hand than stemplots and do not display the actual values observed. For these reasons we prefer stemplots for small data sets. The construction of a histogram is best shown by example. Any statistical software package will of course make a histogram for you, as will your calculator.

Example 1.6 IQ scores Making a histogram

You have probably heard that the distribution of scores on IQ tests follows a bell-shaped pattern. Let's look at some actual IQ scores. Table 1.3 displays the IQ scores of 60 fifth-grade students chosen at random from one school.⁷

Table 1.3 IQ test scores for 60 randomly chosen fifth-grade students

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

Source: James T. Fleming, "The measurement of children's perception of difficulty in reading materials," *Research in the Teaching of English*, 1 (1967), pp. 136–156.

Step 1. Divide the range of the data into classes of equal width. The scores in Table 1.3 range from 81 to 145, so we choose as our classes

$$\begin{aligned} 75 &\leq \text{IQ score} < 85 \\ 85 &\leq \text{IQ score} < 95 \\ &\vdots \\ 145 &\leq \text{IQ score} < 155 \end{aligned}$$

Be sure to specify the classes precisely so that each individual falls into exactly one class. A student with IQ 84 would fall into the first class, but IQ 85 falls into the second.

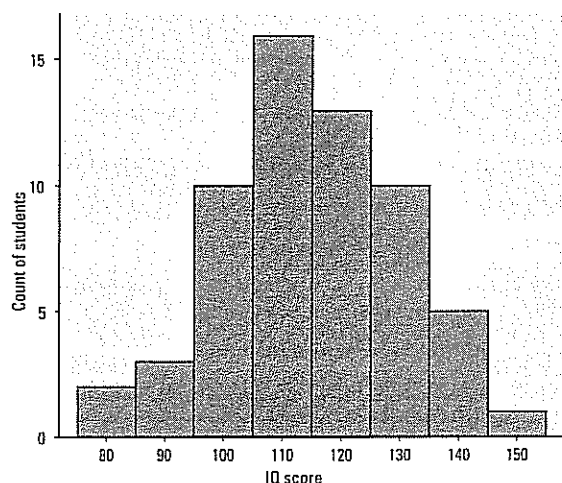
frequency
frequency table **Step 2.** Count the number of individuals in each class. These counts are called *frequencies*, and a table of frequencies for all classes is a *frequency table*.

Class	Count	Class	Count
75 to 84	2	115 to 124	13
85 to 94	3	125 to 134	10
95 to 104	10	135 to 144	5
105 to 114	16	145 to 154	1

Step 3. Draw the histogram and label your axes. First, on the horizontal axis mark the scale for the variable whose distribution you are displaying. That's IQ score. The scale runs from 75 to 155 because that is the span of the classes we chose. The vertical axis contains the scale of counts. Each bar represents a class. The base of the bar covers the class, and the bar height is the class count. There is no horizontal space between the bars unless a class is empty, so that its bar has height zero. Figure 1.6 is our histogram. It does look roughly "bell-shaped."

Figure 1.6

Histogram of the IQ scores of 60 fifth-grade students, for Example 1.6.



A good way to get a feel for how to optimize the presentation of a histogram is to play with an interactive histogram on the computer. The histogram function in the *One-Variable Statistical Calculator* applet on the student CD and Web site is particularly useful because you can change the number of classes by dragging with the mouse. So it's easy to see how the choice of classes affects the histogram.



Activity 1B

The one-variable statistical calculator

The *One-Variable Statistical Calculator* applet on the book's Web site, www.whfreeman.com/tps3e, will make stemplots and histograms. It is intended mainly as a learning tool rather than a replacement for statistical software.

1. Go to the Web site and launch the *One-Variable Statistical Calculator* applet. Your screen should look like this:

Data Sets	Data	Statistics	Histogram	Stemplot
<input type="radio"/> Ages of presidents <input type="radio"/> Fuel economy for 2004 model motor vehicles <input type="radio"/> IQ scores for fifth-graders <input type="radio"/> Newcomb's speed of light <input type="radio"/> Virginia colleges tuition and fees				

2. Choose the “Virginia colleges tuition and fees” data set, and then click on the “Histogram” tab.
 - a. Sketch the default histogram that the applet first presents. If the default graph does not have eight classes, drag it to make a histogram with eight classes and sketch the result.
 - b. Make a histogram with one class and also a histogram with the greatest number of classes that the applet allows. Sketch the results.
 - c. Drag the graph until you find the histogram that you think best pictures the data. How many classes did you choose? Note that if you hold the mouse button down, you can see a popup box that tells you the number of classes. Sketch your final histogram.
 - d. Click on STEMPLOT. Does the applet replicate the stemplot in Figure 1.4 (page 45) in your text? Why do you think the applet won’t replicate the picture in your textbook exactly?
3. Select the data set “IQ scores of fifth graders.”
 - a. See if you can replicate the histogram in Figure 1.6. Drag the scale until your histogram looks as much like the histogram in Figure 1.6 as you can make it. How many classes are there in your histogram? Describe the shape of your histogram.
 - b. Click on STEMPLOT. Sketch the shape of the stemplot. Is the shape of the stemplot similar to the shape of your histogram?
 - c. Select SPLIT STEMS. Does this improve the appearance of the stemplot? In what way?
4. This applet can create a histogram for almost any data set. Select a data set of your choice, click on DATA, and enter your data. Produce a histogram and drag the scale until you are happy with the result. Sketch the histogram. Click on STEMPLOT. Split the stems as needed.

Histogram Tips

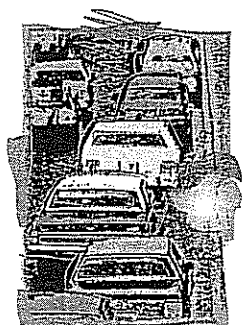
Here are some important properties of histograms to keep in mind when you are constructing a histogram.

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

- Statistical software and graphing calculators will choose the classes for you. The default choice is often a good one, but you can change it if you want. *Beware of letting the device choose the classes.*
- *Use histograms of percents for comparing several distributions with different numbers of observations.* Large sets of data are often reported in the form of frequency tables when it is not practical to publish the individual observations. In addition to the frequency (count) for each class, we may be interested in the fraction or percent of the observations that fall in each class. A histogram of percents looks just like a frequency histogram such as Figure 1.6. Simply relabel the vertical scale to read in percents.

Histograms versus Bar Graphs

Although histograms resemble bar graphs, their details and uses are distinct. A histogram shows the distribution of counts or percents among the values of a single quantitative variable. A bar graph displays the distribution of a categorical variable. The horizontal axis of a bar graph identifies the values of the categorical variable. Draw bar graphs with blank space between the bars to separate the items being compared. Draw histograms with no space, to indicate that all values of the variable are covered.



The vital few? Skewed distributions can show us where to concentrate our efforts. Ten percent of the cars on the road account for half of all carbon dioxide emissions. A histogram of CO₂ emissions would show many cars with small or moderate values and a few with very high values. Cleaning up or replacing these cars would reduce pollution at a cost much lower than that of programs aimed at all cars. Statisticians who work at improving quality in industry make a principle of this: distinguish "the vital few" from "the trivial many."

Examining Distributions

Constructing a graph is only the first step. The next step is to interpret what you see. When you describe a distribution, you should pay attention to the following features.

Examining a Distribution

In any graph of data, look for the **overall pattern** and for striking deviations from that pattern.

You can describe the overall pattern of a distribution by its **shape**, **center**, and **spread**.

An important kind of deviation is an **outlier**, an individual value that falls outside the overall pattern.

We will learn more about describing center and spread numerically in Section 1.2. For now, we can describe the center of a distribution by its *midpoint*, the value with roughly half the observations taking smaller values and half taking larger values. We can describe the spread of a distribution by giving the *smallest and largest values*. Stemplots and histograms display the shape of a distribution in

the same way. Just imagine a stemplot turned on its side so that the larger values lie to the right. Here are some things to look for in describing shape:

- modes*
 - Does the distribution have one or several major peaks, called *modes*? A distribution with one major peak is called *unimodal*.
- unimodal*
- symmetric*
 - Is the distribution approximately symmetric or is it skewed in one direction? A distribution is *symmetric* if the values smaller and larger than its midpoint are mirror images of each other. It is *skewed* to the right if the *right tail* (larger values) is much longer than the *left tail* (smaller values).
- skewed*

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

Example 1.7

IQ scores and Virginia tuitions and fees Interpreting a histogram

What does the histogram of IQ scores in Figure 1.6 (page 50) tell us?

Shape: The distribution is *roughly symmetric* with a *single peak* in the center. We don't expect real data to be perfectly symmetric, so we are satisfied if the two sides of the histogram are roughly similar in shape and extent. **Center:** You can see from the histogram that the midpoint is not far from 110. Looking at the actual data shows that the midpoint is 114. **Spread:** The spread is from about 80 to about 150 (actually 81 to 145). There are no outliers or other strong deviations from the symmetric, unimodal pattern.

The distribution of Virginia's tuitions and fees in Figure 1.4 (page 45) can be summarized as follows:

Shape: The distribution of all two- and four-year colleges is right-skewed. However, if the 23 community college costs are deleted, that is, if we look only at four-year colleges, then the distribution is roughly bell-shaped. **Center:** There are two middle numbers ($n = 60$): 12,901 and 13,150. **Spread:** The data range from 2135 to 34,850. The University of Richmond is a striking deviation. One way to look at this is that 16 students could attend a Virginia community college for about the same cost as that for 1 student at the University of Richmond.

Dealing with Outliers



With small data sets, you can spot outliers by looking for observations that stand apart (either high or low) from the overall pattern of a histogram or stemplot. **Identifying outliers is a matter of judgment.** Look for points that are clearly apart from the body of the data, not just the most extreme observations in a distribution. You should search for an explanation for any outlier. Sometimes outliers point to errors made in recording the data. In other cases, the outlying observation may be caused by equipment failure or other unusual circumstances. In the next section we'll learn a rule of thumb that makes identifying outliers more precise.

Example 1.8**Electronic components**
Outliers

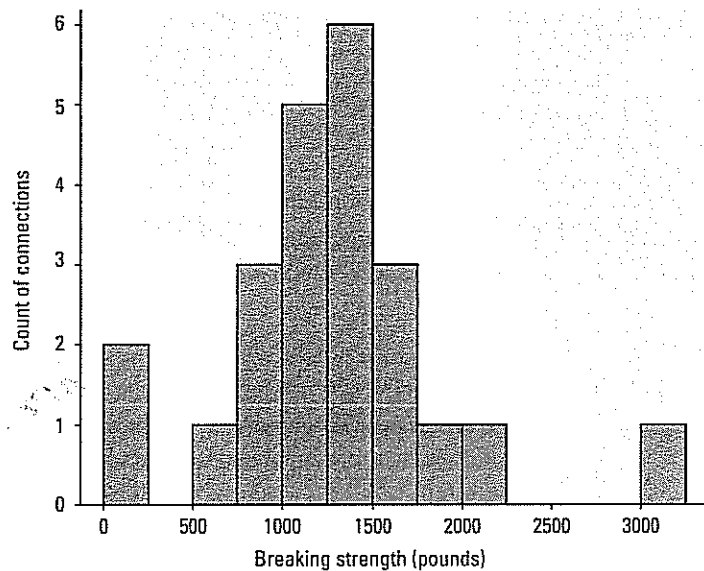
Manufacturing an electronic component requires attaching very fine wires to a semiconductor wafer. If the strength of the bond is weak, the component may fail. Here are measurements on the breaking strength (in pounds) of 23 connections:⁸

0	0	550	750	950	950	1150	1150
1150	1150	1150	1250	1250	1350	1450	1450
1450	1550	1550	1550	1850	2050	3150	

Figure 1.7 is a histogram of these data.

Figure 1.7

Histogram of a distribution with both low and high outliers, for Example 1.8.



We expect the breaking strengths of supposedly identical connections to have a roughly symmetric overall pattern, showing chance variation among the connections. Figure 1.7 does show a symmetric pattern centered at about 1250 pounds—but it also shows three outliers that stand apart from this pattern, two low and one high.

The engineers were able to explain all three outliers. The two low outliers had strength 0 because the bonds between the wire and the wafer were not made. The high outlier at 3150 pounds was a measurement error. Further study of the data can simply omit the three outliers. Note that, in general, it is not a good idea to just delete or ignore outliers. One immediate finding is that the variation in breaking strength is too large—550 pounds to 2050 pounds when we ignore the outliers. The process of bonding wire to wafer must be improved to give more consistent results.

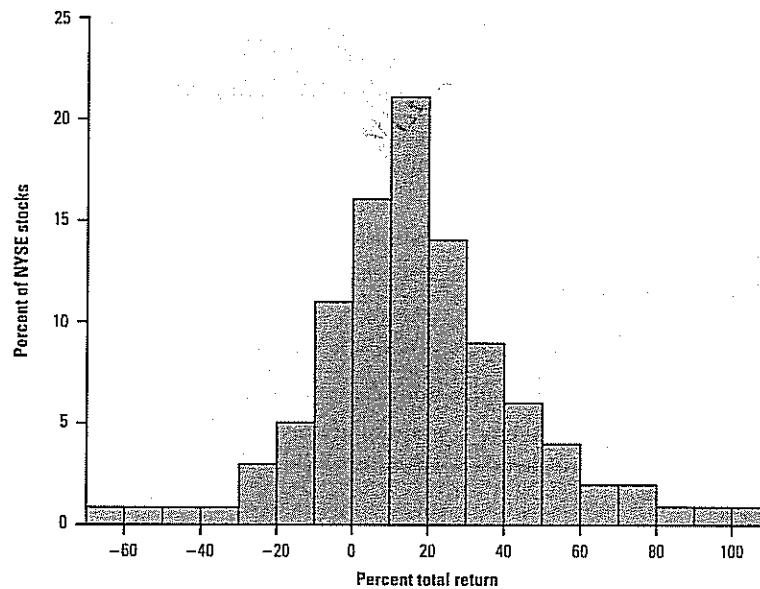


Exercises

1.7 Stock returns The total return on a stock is the change in its market price plus any dividend payments made. Total return is usually expressed as a percent of the beginning price. Figure 1.8 is a histogram of the distribution of total returns for all 1528 stocks listed on the New York Stock Exchange in one year.⁹ Note that it is a histogram of the percents in each class rather than a histogram of counts.

Figure 1.8

Histogram of the distribution of percent total return for all New York Stock Exchange common stocks in one year, for Exercise 1.7.

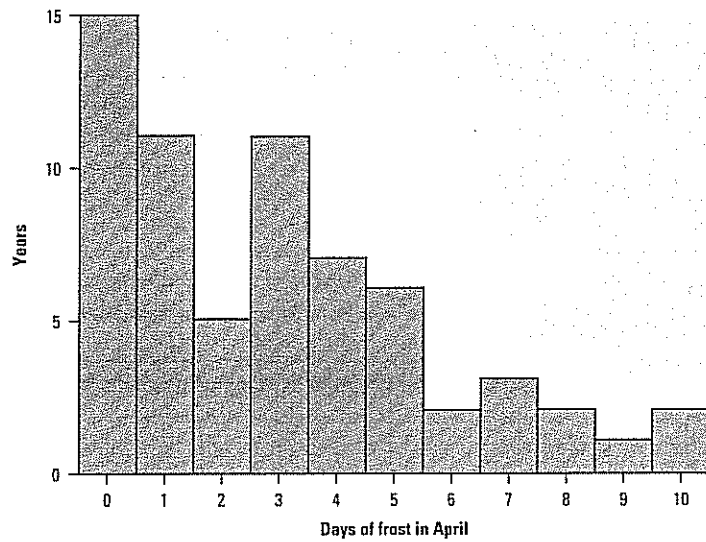


- Describe the overall shape of the distribution of total returns.
- What is the approximate center of this distribution? (For now, take the center to be the value with roughly half the stocks having lower returns and half having higher returns.)
- Approximately what were the smallest and largest total returns? (This describes the spread of the distribution.)
- A return less than zero means that an owner of the stock lost money. About what percent of all stocks lost money?

1.8 Freezing in Greenwich, England Figure 1.9 is a histogram of the number of days in the month of April on which the temperature fell below freezing at Greenwich, England.¹⁰ The data cover a period of 65 years.

Figure 1.9

The distribution of the number of frost days during April at Greenwich, England, over a 65-year period, for Exercise 1.8.

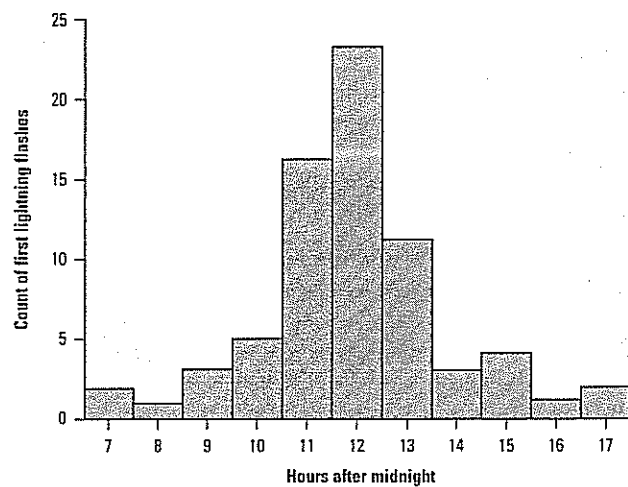


- (a) Describe the shape, center, and spread of this distribution. Are there any outliers?
- (b) In what percent of these 65 years did the temperature never fall below freezing in April?

1.9 Lightning storms Figure 1.10 comes from a study of lightning storms in Colorado. It shows the distribution of the hour of the day during which the first lightning flash for that day occurred. Describe the shape, center, and spread of this distribution.

Figure 1.10

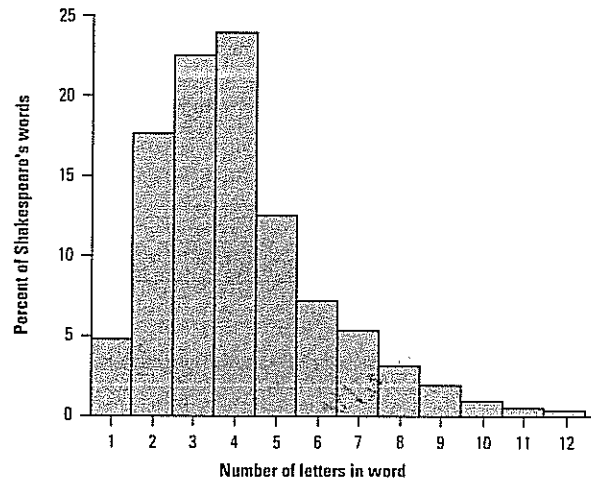
The distribution of the time of the first lightning flash each day at a site in Colorado, for Exercise 1.9.



1.10 Shakespeare Figure 1.11 shows the distribution of lengths of words used in Shakespeare's plays.¹¹ Describe the shape, center, and spread of this distribution.

Figure 1.11

The distribution of lengths of words used in Shakespeare's plays, for Exercise 1.10.



1.11 Presidential ages at inauguration Table 1.4 gives the ages of all U.S. presidents when they took office.

Table 1.4

Ages of the presidents at inauguration

President	Age	President	Age	President	Age
Washington	57	Lincoln	52	Hoover	54
J. Adams	61	A. Johnson	56	F. D. Roosevelt	51
Jefferson	57	Grant	46	Truman	60
Madison	57	Hayes	54	Eisenhower	61
Monroe	58	Garfield	49	Kennedy	43
J. Q. Adams	57	Arthur	51	L. B. Johnson	55
Jackson	61	Cleveland	47	Nixon	56
Van Buren	54	B. Harrison	55	Ford	61
W. H. Harrison	68	Cleveland	55	Carter	52
Tyler	51	McKinley	54	Reagan	69
Polk	49	T. Roosevelt	42	G. H. W. Bush	64
Taylor	64	Taft	51	Clinton	46
Fillmore	50	Wilson	56	G. W. Bush	54
Pierce	48	Harding	55		
Buchanan	65	Coolidge	51		

(a) Make a histogram of the ages of presidents at inauguration. Use class intervals 40 to 44, 45 to 49, and so on. Each interval should contain the left-hand endpoint, but not the right endpoint.

- (b) Describe the shape, center, and spread of the distribution.
- (c) Who was the youngest president? Who was the oldest?
- (d) Was Bill Clinton, at age 46, unusually young?

1.12 Sugar high Carbonated soft drinks are the single biggest source of refined sugars in the American diet.¹² Diets high in refined sugars can promote obesity, which increases the risks of diabetes, high blood pressure, stroke, and heart disease. Sugary soft drinks also promote tooth decay. Forty grams of sugar equates to approximately 10 teaspoons of sugar. The table below shows the number of grams of sugar per 12-fluid-ounce can of 22 popular soft drinks.

Soft drink	Sugar (g)	Soft drink	Sugar (g)
7Up	39	Mello Yellow	47
7Up Plus	2	Minute Maid Orange Soda	48
A&W Root Beer	46	Mountain Dew	46
Cherry Coca-Cola	42	Pepsi-Cola	41
Coca-Cola Classic	39	Pepsi One	0
Coke Zero	0	Pibb Extra	39
Diet Coke	0	Royal Crown Soda	42
Diet Pepsi	0	Sierra Mist	39
Dr Pepper	40	Sprite	38
Fresca	0	Sunkist Orange Soda	52
IBC Root Beer	43	Welch's Sparkling Grape Soda	51

- (a) Construct an appropriate graph of these data. A dotplot might suffice, but a stemplot or histogram might be preferable due to the large spread. If you use the *One-Variable Statistical Calculator* applet, you can experiment with different class widths.
- (b) Describe what you see; that is, describe the distribution.



Technology Toolbox



Making calculator histograms

- Enter the presidential age data from Exercise 1.11 (page 57) in your statistics list editor.

TI-83/84

- Press **STAT** and choose 1:Edit. . . .
- Type the values into list L₁.

L1	L2	L3	1
57	---	---	
61			
57			
57			
58			
57			
61			
L1={57,61,57,57...			

TI-89

- Press **APPS**, choose 1:FlashApps, then select Stats/List Editor and press **ENTER**.
- Type the values into list1.

list1	list2	list3	list4
57	---	---	---
61			
57			
57			
58			
57			
list1[1]=57			
MAIN RAD AUTO FUNC 1/6			

- Set up a histogram in the statistics plots menu.

- Press **2nd** **Y=** (STAT PLOT).
- Press **ENTER** to go into Plot1.
- Adjust your settings as shown.

Plot1	Plot2	Plot3
On	Off	
Type:		
Xlist:	L1	
Freq:	1	

- Press **F2** and choose 1:Plot Setup. . . .
- With Plot 1 highlighted, press **F1** to define.
- Change Hist. Bucket Width to 5, as shown.

Define Plot 1	
Plot Type	Histogram→
BarX	1/VA
Y	list1
Hist.Bucket Width	5
Use Freq and Categories?	NO→
Freq	
Category	
Include Categories	0:
Enter=OK	ESC=CANCEL
USE ← AND → TO OPEN CHOICES	

- Set the window to match the class intervals chosen in Exercise 1.11.

- Press **WINDOW**.
- Enter the values shown.

WINDOW
Xmin=35
Xmax=75
Xscl=5
Ymin=-3
Ymax=15
Yscl=1
Xres=1

- Press **2nd** **F2** (WINDOW).
- Enter the values shown.

FIX	FIX
Tools	Tools
Zoom	
xmin=35.	
xmax=75.	
xscl=5.	
ymin=-3.	
ymax=15.	
yscl=1.	
xres=1.	
MAIN DEG AUTO FUNC	

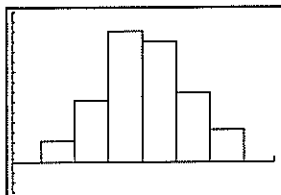
Technology Toolbox



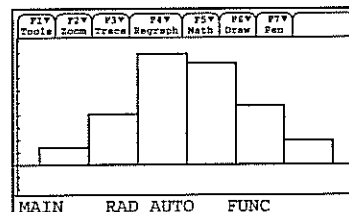
Making calculator histograms (continued)

4. Graph the histogram.

- Press **GRAPH**.



- Press **2ND** **GRAPH** (GRAPH).



5. Save the data in a named list for later use.

- From the home screen, type the command $L1 \rightarrow \text{PREZ}$ (list1 \rightarrow prez on the TI-89) and press **ENTER**. The data are now stored in a list called PREZ.

```
L1→PREZ
{ 57 61 57 57 58 ...
```

```
F1V F2V F3V F4V F5 F6V
Tools Algebra Calc Other Programs Clean Up

■list1→prez
{ 57 61 57 57 58 57 ▶
list1→prez
MAIN DEG AUTO FUNC 1/30
```

Relative Frequency and Cumulative Frequency

A histogram does a good job of displaying the distribution of values of a quantitative variable. But it tells us little about the relative standing of an individual observation. If we want this type of information, we should construct a relative cumulative frequency graph, often called an ogive (pronounced O-JIVE).

Example 1.9

Presidents

Constructing a relative cumulative frequency graph (ogive)

In Exercise 1.11, you were asked to make a histogram of the ages of U.S. presidents when they were inaugurated. Now we will examine where some specific presidents fall within the age distribution.

How to construct an ogive (relative cumulative frequency graph):

Step 1: Decide on class intervals and make a frequency table, just as in making a histogram. Add three columns to your frequency table: relative frequency, cumulative frequency, and relative cumulative frequency.

- To get the values in the *relative frequency* column, divide the count in each class interval by 43, the total number of presidents. Multiply by 100 to convert to a percent.

- To fill in the *cumulative frequency* column, add the counts in the frequency column that fall in or below the current class interval.
- For the *relative cumulative frequency* column, divide the entries in the cumulative frequency column by 43, the total number of individuals.

Here is the frequency table with the relative frequency, cumulative frequency, and relative cumulative frequency columns added.

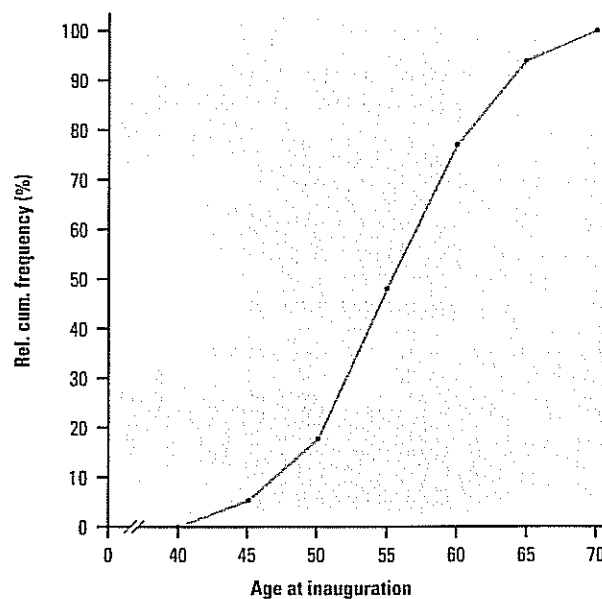
Class	Frequency	Relative frequency	Cumulative frequency	Relative cumulative frequency
40–44	2	$\frac{2}{43} = 0.047$, or 4.7%	2	$\frac{2}{43} = 0.047$, or 4.7%
45–49	6	$\frac{6}{43} = 0.140$, or 14.0%	8	$\frac{8}{43} = 0.186$, or 18.6%
50–54	13	$\frac{13}{43} = 0.302$, or 30.2%	21	$\frac{21}{43} = 0.488$, or 48.8%
55–59	12	$\frac{12}{43} = 0.279$, or 27.9%	33	$\frac{33}{43} = 0.767$, or 76.7%
60–64	7	$\frac{7}{43} = 0.163$, or 16.3%	40	$\frac{40}{43} = 0.930$, or 93.0%
65–69	3	$\frac{3}{43} = 0.070$, or 7.0%	43	$\frac{43}{43} = 1.000$, or 100%
Total	43			

Step 2: Label and scale your axes and title your graph. Label the horizontal axis “Age at inauguration” and the vertical axis “Relative cumulative frequency.” Scale the horizontal axis according to your choice of class intervals and the vertical axis from 0% to 100%.

Step 3: Plot a point corresponding to the relative cumulative frequency in each class interval at the *left endpoint* of the *next* class interval. For example, for the 40 to 44 interval, plot a point at a height of 4.7% above the age value of 45. This means that 4.7% of presidents were inaugurated before they were 45 years old. Begin your ogive with a point at a height of 0% at the left endpoint of the lowest class interval. Connect consecutive points with a line segment to form the ogive. The last point you plot should be at a height of 100%. Figure 1.12 shows the completed ogive.

Figure 1.12

Relative cumulative frequency plot (ogive) for the ages of U.S. presidents at inauguration.



How to locate an individual within the distribution:

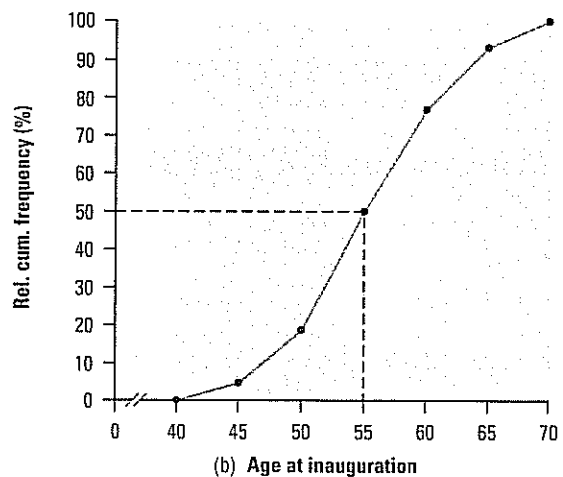
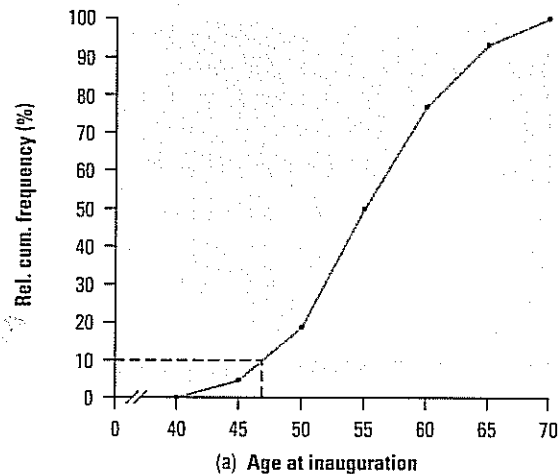
What about Bill Clinton? He was age 46 when he took office. To find his relative standing, draw a vertical line up from his age (46) on the horizontal axis until it meets the ogive. Then draw a horizontal line from this point of intersection to the vertical axis. Based on Figure 1.13(a), we would estimate that Bill Clinton's age places him at the 10% *relative cumulative frequency* mark. That tells us that about 10% of all U.S. presidents were the same age as or younger than Bill Clinton when they were inaugurated. Put another way, President Clinton was younger than about 90% of all U.S. presidents based on his inauguration age.

How to locate a value corresponding to a percentile:

What is the center of the distribution? To answer this question, draw a horizontal line across from the vertical axis at a height of 50% until it meets the ogive. From the point of intersection, draw a vertical line down to the horizontal axis. In Figure 1.13(b), the value on the horizontal axis is about 55. So about 50% of all presidents were 55 years old or younger when they took office, and 50% were older. The center of the distribution is 55.

Figure 1.13

Ogives of presidents' ages at inauguration are used to (a) locate President Clinton within the distribution and (b) determine the 50th percentile, which is age 55.



Time Plots



Whenever data are collected over time, it is a good idea to plot the observations in time order. Displays of the distribution of a variable that ignore time order, such as stemplots and histograms, can be misleading when there is systematic change over time.

Time Plot

A time plot of a variable plots each observation against the time at which it was measured. Always put time on the horizontal scale of your plot and the variable you are measuring on the vertical scale. Connecting the data points by lines helps emphasize any change over time.

Example 1.10 Gas prices

Time plot with seasonal variation

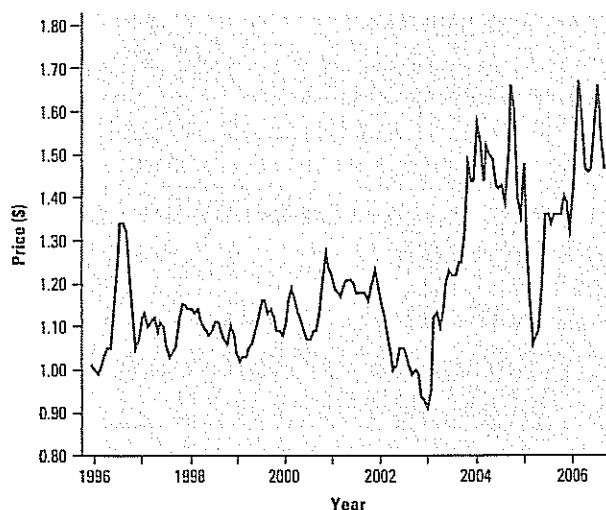
Figure 1.14 is a time plot of the average retail price of regular gasoline each month for the years 1996 to 2006.¹³ You can see a drop in prices in 1998 when an economic crisis in Asia reduced demand for fuel. You can see rapid price increases in 2000 and 2003 due to instability in the Middle East and OPEC production limits. These deviations are so large that overall patterns are hard to see. Since 2002 there has been a generally upward trend in gas prices. In the second half of 2005, after Hurricane Katrina disrupted the production and flow of oil from the Gulf of Mexico, gas prices peaked near \$3.00 per gallon.

There is nonetheless a clear *trend* of increasing price. Much of this trend just reflects inflation, the rise in the overall price level during these years. In addition, a close look at the plot shows *seasonal variation*, a regular rise and fall that recurs each year. Americans drive more in the summer vacation season, so the price of gasoline rises each spring, then drops in the fall as demand goes down.

seasonal
variation

Figure 1.14

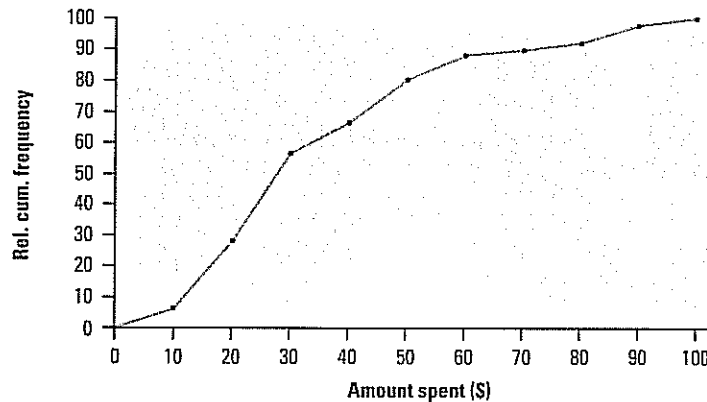
Time plot of the average monthly price of regular gasoline from 1996 to 2006, for Example 1.10.



Exercises

1.13 Shopping spree, II Figure 1.15 is an ogive of the amount spent by grocery shoppers in Exercise 1.6 (page 48).

Figure 1.15 Ogive of amount spent by grocery shoppers, for Exercise 1.13.



- Estimate the center of this distribution. Explain your method.
- What is the relative cumulative frequency for the shopper who spent \$17.00?
- Draw the histogram that corresponds to the ogive.

1.14 Glucose levels People with diabetes must monitor and control their blood glucose level. The goal is to maintain “fasting plasma glucose” between about 90 and 130 milligrams per deciliter (mg/dl) of blood. Here are the fasting plasma glucose levels for 18 diabetics enrolled in a diabetes control class, five months after the end of the class:¹⁴

141	158	112	153	134	95	96	78	148
172	200	271	103	172	359	145	147	255

- Make a stemplot of these data and describe the main features of the distribution. (You will want to round and also split stems.) Are there outliers? How well is the group as a whole achieving the goal for controlling glucose levels?
- Construct a relative cumulative frequency graph (ogive) for these data.
- Use your graph from part (b) to answer the following questions:
 - What percent of blood glucose levels were between 90 and 130?
 - What is the center of this distribution?
 - What relative cumulative frequency is associated with a blood glucose level of 130?



1.15 Birthrates The table below shows the number of births in the United States and the birthrates at 10-year intervals from 1960 to 2000. The birthrate is the number of births per 1000 population.¹⁵

Year	Total number	Rate
1960	4,257,850	23.7
1970	3,731,386	18.4
1980	3,612,258	15.9
1990	4,092,994	16.7
2000	4,058,814	14.4

- Construct a time plot for the birthrate, 1960 to 2000.
- Is there a trend in the birthrate? If so, describe the trend in a sentence or two.
- List some factors that you think might explain what you see in your birthrate time plot.
- Construct a time plot for the total number of births.
- Describe what is happening over time for the total number of births.
- Briefly explain how you can have such different plots for the two variables.

1.16 Life expectancy Most people are aware that life expectancy, the number of years a person can expect to live, is much longer now than it was, say, a century ago. Here are the numbers for women provided by the National Center for Health Statistics.

Year	Life expectancy (female)	Year	Life expectancy (female)
1900	48.3	1960	73.1
1910	51.8	1970	74.7
1920	54.6	1980	77.5
1930	61.6	1990	78.8
1940	65.2	2000	79.5
1950	71.1		

- Construct a time plot for these data.
- Describe what you see about the life expectancy of females over the last hundred years.

1.17 The speed of light Light travels fast, but it is not transmitted instantaneously. Light takes over a second to reach us from the moon and over 10 billion years to reach us from the most distant objects observed so far in the expanding universe. Because radio and radar also travel at the speed of light, an accurate value for that speed is important in communicating with astronauts and orbiting satellites. An accurate value for the speed of light is also important to computer designers because electrical signals travel at light speed. The first reasonably accurate measurements of the speed of light were made over a hundred years ago by A. A. Michelson and Simon Newcomb. Table 1.5 contains 66 measurements made by Newcomb between July and September 1882.

Table 1.5 Newcomb's measurements of the passage time of light

28	26	33	24	34	-44	27	16	40	-2	29	22	24	21
25	30	23	29	31	19	24	20	36	32	36	28	25	21
28	29	37	25	28	26	30	32	36	26	30	22	36	23
27	27	28	27	31	27	26	33	26	32	32	24	39	28
24	25	32	25	29	27	28	29	16	23				

Source: S. M. Stigler, "Do robust estimators work with real data?" *Annals of Statistics*, 5 (1977), pp. 1055-1078.

Newcomb measured the time in seconds that a light signal took to pass from his laboratory on the Potomac River to a mirror at the base of the Washington Monument and back, a total distance of about 7400 meters. Just as you can compute the speed of a car from the time required to drive a mile, Newcomb could compute the speed of light from the passage time. Newcomb's first measurement of the passage time of light was 0.000024828 second, or 24,828 nanoseconds. (There are 10^9 nanoseconds in a second.) The entries in Table 1.5 record only the deviation from 24,800 nanoseconds.

- Construct an appropriate graphical display for these data. Justify your choice of graph.
 - Describe the distribution of Newcomb's speed of light measurements.
 - Make a time plot of Newcomb's values. They are listed in order from left to right, starting with the top row.
 - What does the time plot tell you that the graph you made in part (a) does not?
- Lesson:* Sometimes you need to make more than one graphical display to uncover all of the important features of a distribution.

1.18 Civil unrest The years around 1970 brought unrest to many U.S. cities. Here are data on the number of civil disturbances in each three-month period during the years 1968 to 1972:

Period			Count	Period			Count
1968	Jan.-Mar.		6	1970	July-Sept.		20
	Apr.-June		46		Oct.-Dec.		6
	July-Sept.		25	1971	Jan.-Mar.		12
	Oct.-Dec.		3		Apr.-June		21
1969	Jan.-Mar.		5	1972	July-Sept.		5
	Apr.-June		27		Oct.-Dec.		1
	July-Sept.		19		Jan.-Mar.		3
	Oct.-Dec.		6		Apr.-June		8
1970	Jan.-Mar.		26		July-Sept.		5
	Apr.-June		24		Oct.-Dec.		5

- Make a time plot of these counts. Connect the points in your plot by straight-line segments to make the pattern clearer.
- Describe the trend and the seasonal variation in this time series. Can you suggest an explanation for the seasonal variation in civil disorders?

Section 1.1 Summary

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

To describe a distribution, begin with a graph. **Bar graphs** and **pie charts** display the distributions of categorical variables. These graphs use the counts or percents of the categories. **Stemplots** and **histograms** display the distributions of quantitative variables. Stemplots separate each observation into a stem and a one-digit leaf. Histograms plot the **frequencies** (counts) or percents of equal-width classes of values.

When examining a distribution, look for **shape**, **center**, and **spread**, and for clear **deviations** from the overall shape. Some distributions have simple shapes, such as **symmetric** or **skewed**. The number of **modes** (major peaks) is another aspect of overall shape. Not all distributions have a simple overall shape, especially when there are few observations.

Outliers are observations that lie outside the overall pattern of a distribution. Always look for outliers and try to explain them.

A **relative cumulative frequency graph** (also called an **ogive**) is a good way to see the relative standing of an observation.

When observations on a variable are taken over time, make a **time plot** that graphs time horizontally and the values of the variable vertically. A time plot can reveal **trends** or other changes over time.

Section 1.1 Exercises

1.19 Ranking colleges Popular magazines rank colleges and universities on their “academic quality” in serving undergraduate students. Describe five variables that you would like to see measured for each college if you were choosing where to study. Give reasons for each of your choices.

1.20 Shopping spree, III Enter the amount of money spent in a supermarket from Exercise 1.6 (page 48) into your calculator. Then use the information in the Technology Toolbox (page 59) to construct a histogram. Use ZoomStat/ZoomData first to see what the calculator chooses for class widths. Then, in the calculator’s WINDOW, choose new settings that are more sensible. Compare your histogram with the stemplots you made in Exercise 1.6. List at least one advantage that each plot has that the other plots don’t have.

1.21 College costs The Department of Education estimates the “average unmet need” for undergraduate students—the cost of school minus estimated family contributions and financial aid. Here are the averages for full-time students at four types of institution in the 1999–2000 academic year:¹⁶

Public 2-year	Public 4-year	Private nonprofit 4-year	Private for profit
\$4495	\$4818	\$8257	\$8296

Make a bar graph of these data. Write a one-sentence conclusion about the unmet needs of students. Explain clearly why it is incorrect to make a pie chart.

1.22 New-vehicle survey The J. D. Power Initial Quality Study polls more than 50,000 buyers of new motor vehicles 90 days after their purchase. A two-page questionnaire asks about “things gone wrong.” Here are data on problems per 100 vehicles for vehicles made by Toyota and by General Motors in recent years. Toyota has been the industry leader in quality. Make two time plots in the same graph to compare Toyota and GM. What are the most important conclusions you can draw from your graph?

	1998	1999	2000	2001	2002	2003	2004
GM	187	179	164	147	130	134	120
Toyota	156	134	116	115	107	115	101

1.23 Senior citizens, I The population of the United States is aging, though less rapidly than in other developed countries. Here is a stemplot of the percents of residents aged 65 and over in the 50 states, according to the 2000 census. The stems are whole percents and the leaves are tenths of a percent.

5	7
6	
7	
8	5
9	6 7 9
10	6
11	0 2 2 3 3 6 7 7
12	0 0 1 1 1 1 3 4 4 5 7 8 9
13	0 0 0 1 2 2 3 3 3 4 5 5 6 8
14	0 3 4 5 7 9
15	3 6
16	
17	6

(a) There are two outliers: Alaska has the lowest percent of older residents, and Florida has the highest. What are the percents for these two states?

(b) Ignoring Alaska and Florida, describe the shape, center, and spread of this distribution.

1.24 Senior citizens, II Make another stemplot of the percent of residents aged 65 and over in the states other than Alaska and Florida by splitting stems in the plot from the previous exercise. Which plot do you prefer? Why?

1.25 The statistics of writing style Numerical data can distinguish different types of writing, and sometimes even individual authors. Here are data on the percent of words of 1 to 15 letters used in articles in *Popular Science* magazine:¹⁷

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

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

(b) How does the distribution of lengths of words used in *Popular Science* compare with the similar distribution in Figure 1.11 (page 57) for Shakespeare’s plays? Look in particular at short words (2, 3, and 4 letters) and very long words (more than 10 letters).



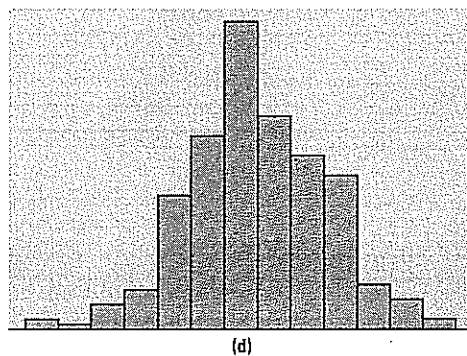
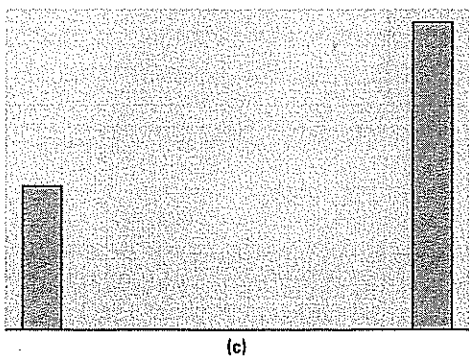
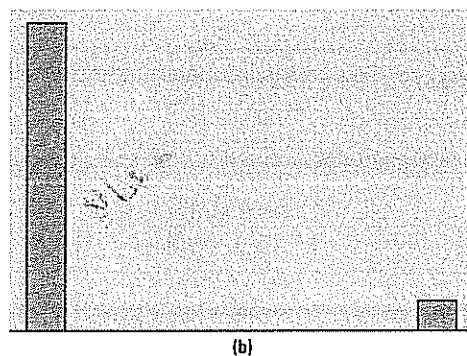
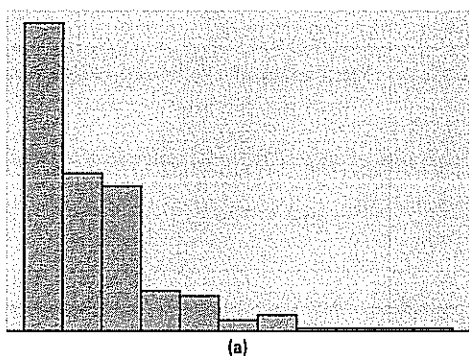
1.26 Student survey A survey of a large high school class asked the following questions:

1. Are you female or male? (In the data, male = 0, female = 1.)
2. Are you right-handed or left-handed? (In the data, right = 0, left = 1.)
3. What is your height in inches?
4. How many minutes do you study on a typical weeknight?

Figure 1.16 shows histograms of the student responses, in scrambled order and without scale markings. Which histogram goes with each variable? Explain your reasoning.

Figure 1.16

Match each histogram with its variable, for Exercise 1.26.



1.2 Describing Distributions with Numbers

Interested in a sporty car? Worried that it might use too much gas? The Environmental Protection Agency lists most such vehicles in its “two-seater” or “mini-compact” categories. Table 1.6 on the next page gives the city and highway gas mileage for cars in these groups. (The mileages are for the basic engine and transmission combination for each car.) We want to compare two-seaters with mini-compacts and city mileage with highway mileage. We can begin with graphs, but numerical summaries make the comparisons more specific.

Table 1.6 Fuel economy (miles per gallon) for 2004 model motor vehicles

Two-seater Cars			Minicompact Cars		
Model	City	Highway	Model	City	Highway
Acura NSX	17	24	Aston Martin Vanquish	12	19
Audi TT Roadster	20	28	Audi TT Coupe	21	29
BMW Z4 Roadster	20	28	BMW 325CI	19	27
Cadillac XLR	17	25	BMW 330CI	19	28
Chevrolet Corvette	18	25	BMW M3	16	23
Dodge Viper	12	20	Jaguar XK8	18	26
Ferrari 360 Modena	11	16	Jaguar XKR	16	23
Ferrari Maranello	10	16	Lexus SC 430	18	23
Ford Thunderbird	17	23	Mini Cooper	25	32
Honda Insight	60	66	Mitsubishi Eclipse	23	31
Lamborghini Gallardo	9	15	Mitsubishi Spyder	20	29
Lamborghini Murcielago	9	13	Porsche Cabriolet	18	26
Lotus Esprit	15	22	Porsche Turbo 911	14	22
Maserati Spyder	12	17			
Mazda Miata	22	28			
Mercedes-Benz SL500	16	23			
Mercedes-Benz SL600	13	19			
Nissan 350Z	20	26			
Porsche Boxster	20	29			
Porsche Carrera 911	15	23			
Toyota MR2	26	32			

Source: U.S. Environmental Protection Agency, "Model Year 2004 Fuel Economy Guide," found online at www.fueleconomy.gov.

A brief description of a distribution should include its *shape* and numbers describing its *center* and *spread*. We describe the shape of a distribution based on inspection of a histogram or a stemplot. Now you will learn specific ways to use numbers to measure the center and spread of a distribution. You can calculate these numerical measures for any quantitative variable. But to interpret measures of center and spread, and to choose among the several measures you will learn, you must think about the shape of the distribution and the meaning of the data. The numbers, like graphs, are aids to understanding, not "the answer" in themselves.

Measuring Center: The Mean

Numerical description of a distribution begins with a measure of its center or average. The two common measures of center are the *mean* and the *median*. The mean is the "average value," and the median is the "middle value." These are two different ideas for "center," and the two measures behave differently. We need precise rules for the mean and the median.

The Mean \bar{x}

To find the **mean** \bar{x} of a set of observations, add their values and divide by the number of observations. If the n observations are x_1, x_2, \dots, x_n , their mean is

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n}$$

or, in more compact notation,

$$\bar{x} = \frac{\sum x_i}{n}$$

The Σ (capital Greek sigma) in the formula for the mean is short for “add them all up.” The bar over the x indicates the mean of all the x -values. Pronounce the mean \bar{x} as “x-bar.” This notation is so common that writers who are discussing data use \bar{x} , \bar{y} , etc. without additional explanation. The subscripts on the observations x_i are just a way of keeping the n observations distinct. They do not necessarily indicate order or any other special facts about the data.

Example 1.11 *Mean highway mileage for two-seaters*
 Calculating \bar{x}

The mean highway mileage for the 21 two-seaters in Table 1.6 is

$$\begin{aligned}\bar{x} &= \frac{x_1 + x_2 + \dots + x_n}{n} \\ &= \frac{24 + 28 + 28 \dots + 32}{21} \\ &= \frac{518}{21} = 24.7 \text{ miles per gallon}\end{aligned}$$

In practice, you can key the data into your calculator and request 1-Var Stats, or into software and request descriptive statistics.



resistant measure

The data for Example 1.11 contain an outlier: the Honda Insight is a hybrid gas-electric car that doesn't belong in the same category as the 20 gasoline-powered two-seater cars. If we exclude the Insight, the mean highway mileage drops to 22.6 mpg. The single outlier adds more than 2 mpg to the mean highway mileage. This illustrates an important weakness of the mean as a measure of center: **the mean is sensitive to the influence of a few extreme observations.** These may be outliers, but a skewed distribution that has no outliers will also pull the mean toward its long tail. Because the mean cannot resist the influence of extreme observations, we say that it is not a **resistant measure** of center. A measure that is resistant does more than limit the influence of outliers. Its value does not respond strongly to changes in a few observations, no matter how large those changes may be. The mean fails this requirement because we can make the mean as large as we wish by making a large enough increase in just one observation.

Measuring Center: The Median

We used the midpoint of a distribution as an informal measure of center in the previous section. The *median* is the formal version of the midpoint, with a specific rule for calculation.

The Median M

The **median M** is the midpoint of a distribution, the number such that half the observations are smaller and the other half are larger. To find the median of a distribution:

1. Arrange all observations in order of size, from smallest to largest.
2. If the number of observations n is odd, the median M is the center observation in the ordered list. Find the *location* of the median by counting $(n + 1)/2$ observations up from the bottom of the list.
3. If the number of observations n is even, the median M is the average of the two center observations in the ordered list. The location of the median is again $(n + 1)/2$ from the bottom of the list.

Note that the formula $(n + 1)/2$ does *not* give the median, just the location of the median in the ordered list. Medians require little arithmetic, so they are easy to find by hand for small sets of data. Arranging even a moderate number of observations in order is tedious, however, so that finding the median by hand for larger sets of data is unpleasant. Even simple calculators have an \bar{x} button, but you will need computer software or a graphing calculator to automate finding the median.

Example 1.12

Median highway mileage for two-seaters Finding the median by hand

To find the median highway mileage for 2004 model two-seater cars, arrange the data in increasing order:

13 15 16 16 17 19 20 22 23 23 23 24 25 25 26 28 28 28 29 32 66

Be sure to list *all* observations, even if they repeat the same value. The median is the bold 23, the 11th observation in the ordered list. You can find the median by eye—there are 10 observations to the left and 10 to the right. Or you can use the rule $(n + 1)/2 = 22/2 = 11$ to locate the median in the list.

What happens if we drop the Honda Insight? The remaining 20 cars have highway mileages

13 15 16 16 17 19 20 22 23 23 23 24 25 25 26 28 28 28 29 32

Because the number of observations $n = 20$ is even, there is no center observation. There is a center pair—the bold pair of 23s have 9 observations to their left and 9 to their right. The median M is the average of the center pair, which is 23. The rule $(n + 1)/2 = 21/2 = 10.5$ for the position of the median in the list says that the median is at location “ten and one-half,” that is, halfway between the 10th and 11th observations.

You see that the median is more resistant than the mean. Removing the Honda Insight did not change the median at all. Even if we mistakenly enter the Insight's mileage as 660 rather than 66, the median remains 23. The very high value is simply one observation to the right of center.



Activity 1C

The Mean and Median applet

The *Mean and Median* applet on the book's Web site, www.whfreeman.com/tps3e, allows you to place observations on a line and see their mean and median visually.

1. Go to the Web site and launch the *Mean and Median* applet. The applet consists of a horizontal line and a trash can.
2. Place two observations on the line by clicking below it. Why does only one arrow appear?
3. Place three observations on the line by clicking below it, two close together near the center of the line and one somewhat to the right of these two.
 - a. Pull the single rightmost observation out to the right. (Place the cursor on the point, hold down the mouse button, and drag the point.) How does the mean behave? How does the median behave? Explain briefly why each measure acts as it does.
 - b. Now drag the rightmost point to the left as far as you can. What happens to the mean? What happens to the median as you drag this point past the other two? (Watch carefully.)
4. Place five observations on the line by clicking below it.
 - a. Add one additional observation *without changing the median*. Where is your new point?
 - b. Use the applet to convince yourself that when you add yet another observation (there are now seven in all), the median does not change no matter where you put the seventh point. Explain why this must be true.

Mean versus Median

The median and mean are the most common measures of the center of a distribution. The mean and median of a symmetric distribution are close together. If the distribution is exactly symmetric, the mean and median are exactly the same. In a skewed distribution, the mean is farther out in the long tail than is the median. For example, the distribution of the sizes of the endowments of colleges and



universities is strongly skewed to the right. Most institutions have modest endowments, but a few are very wealthy. The median endowment of colleges and universities in 2003 was \$70 million—but the mean endowment was over \$320 million. The few wealthy institutions pulled the mean up but did not affect the median. Don't confuse the "average" value of a variable (the mean) with its "typical" value, which we might describe by the median.

We can now give a better answer to the question of how to deal with outliers in data. First, look at the data to identify outliers and investigate their causes. You can then correct outliers if they are wrongly recorded, delete them for good reason, or otherwise give them individual attention. If you are interested only in Virginia four-year colleges, for example, then it makes sense to delete the 23 community colleges from consideration. Then among Virginia four-year colleges, the University of Richmond's \$34,850 figure is now an outlier, as you will see shortly. It would be inappropriate to delete this high outlier. If you have no clear reason to drop outliers, you may want to use resistant methods, so that outliers have little influence over your conclusions. The choice is often a matter of judgment. The government's fuel economy guide lists the Honda Insight with the other two-seaters in Table 1.6. We might choose to report median rather than mean gas mileage for all two-seaters to avoid giving too much influence to one car model. In fact, we think that the Insight doesn't belong, so we will omit it from further analysis of these data.

Exercises

1.27 Quiz grades Joey's first 14 quiz grades in a marking period were

86 84 91 75 78 80 74 87 76 96 82 90 98 93

- Use the formula to calculate the mean. Check using "1-Var Stats" on your calculator.
- Suppose Joey has an unexcused absence for the 15th quiz, and he receives a score of zero. Determine his final quiz average. What property of the mean does this situation illustrate? Write a sentence about the effect of the zero on Joey's quiz average that mentions this property.
- What kind of plot would best show Joey's distribution of grades? Assume an eight-point grading scale (A: 93 to 100; B: 85 to 92; etc.). Make an appropriate plot, and be prepared to justify your choice.

1.28 SSHA scores, I The Survey of Study Habits and Attitudes (SSHA) is a psychological test that evaluates college students' motivation, study habits, and attitudes toward school. A private college gives the SSHA to a sample of 18 of its incoming first-year women students. Their scores are

154 109 137 115 152 140 154 178 101
103 126 126 137 165 165 129 200 148

- Make a stemplot of these data. The overall shape of the distribution is irregular, as often happens when only a few observations are available. Are there any potential outliers? About where is the median of the distribution (the score with half the scores above it and half below)? What is the spread of the scores (ignoring any outliers)?

(b) Find the mean score from the formula for the mean. Then enter the data into your calculator. You can find the mean from the home screen as follows:

TI-83/84

- Press **2nd** **STAT** (LIST) **►►** (MATH).
- Choose 3: mean (, enter list name, press **ENTER**.

TI-89

- Press **CATALOG** then **5** (M).
- Choose mean (, type list name, press **ENTER**.

(c) Find the median of these scores. Which is larger: the median or the mean? Explain why.

1.29 Baseball player salaries Suppose a Major League Baseball team's mean yearly salary for a player is \$1.2 million, and that the team has 25 players on its active roster. What is the team's annual payroll for players? If you knew only the median salary, would you be able to answer the question? Why or why not?

1.30 Mean salary? Last year a small accounting firm paid each of its five clerks \$22,000, two junior accountants \$50,000 each, and the firm's owner \$270,000. What is the mean salary paid at this firm? How many of the employees earn less than the mean? What is the median salary? Write a sentence to describe how an unethical recruiter could use statistics to mislead prospective employees.

1.31 U.S. incomes The distribution of household incomes in the United States is strongly skewed to the right. In 2003, the mean and median household incomes in America were \$43,318 and \$59,067. Which of these numbers is the mean and which is the median? Explain your reasoning.

1.32 Home run records Who is baseball's greatest home run hitter? In the summer of 1998, Mark McGwire and Sammy Sosa captured the public's imagination with their pursuit of baseball's single-season home run record (held by Roger Maris). McGwire eventually set a new standard with 70 home runs. Barry Bonds broke McGwire's record when he hit 73 home runs in the 2001 season. How does this accomplishment fit Bonds's career? Here are Bonds's home run counts for the years 1986 (his rookie year) to 2004:

1986	1987	1988	1989	1990	1991	1992	1993	1994	
16	25	24	19	33	25	34	46	37	
1995	1996	1997	1998	1999	2000	2001	2002	2003	2004
33	42	40	37	34	49	73	46	45	45

(a) Calculate the mean and median of Bonds's home run data. What do these two numbers tell you about the distribution of the data?

(b) Make a stemplot of these data.

(c) What would you say is Bonds's home run production for a typical year (1986 to 2004)? Explain your reasoning in a sentence or two.

Measuring Spread: The Quartiles

A measure of center alone can be misleading. Two nations with the same median family income are very different if one has extremes of wealth and poverty and the

other has little variation among families. A drug with the correct mean concentration of active ingredient is dangerous if some batches are much too high and others much too low. We are interested in the *spread* or *variability* of incomes and drug potencies as well as their centers. *The simplest useful numerical description of a distribution consists of both a measure of center and a measure of spread.*

range

One way to measure spread is to calculate the *range*, which is the difference between the largest and smallest observations. For example, the number of home runs Barry Bonds has hit in a season has a *range* of $73 - 16 = 57$. The range shows the full spread of the data. But it depends on only the smallest observation and the largest observation, which may be outliers.

*p*th percentile

We can describe the spread or variability of a distribution by giving several percentiles. The *p*th percentile of a distribution is the value such that *p* percent of the observations fall at or below it. The median is just the 50th percentile, so the use of percentiles to report spread is particularly appropriate when the median is our measure of center. The most commonly used percentiles other than the median are the *quartiles*. The first quartile is the 25th percentile, and the third quartile is the 75th percentile. (The second quartile is the median itself.) To calculate a percentile, arrange the observations in increasing order and count up the required percent from the bottom of the list. Our definition of percentiles is a bit inexact, because there is not always a value with exactly *p* percent of the data at or below it. We will be content to take the nearest observation for most percentiles, but the quartiles are important enough to require an exact rule.

The Quartiles Q_1 and Q_3

To calculate the *quartiles*:

1. Arrange the observations in increasing order and locate the median M in the ordered list of observations.
2. The first quartile Q_1 is the median of the observations whose position in the ordered list is to the left of the location of the overall median.
3. The third quartile Q_3 is the median of the observations whose position in the ordered list is to the right of the location of the overall median.

Here is an example that shows how the rules for the quartiles work for both odd and even numbers of observations.

Example 1.13

Highway mileage Calculating quartiles

The highway mileages of the 20 gasoline-powered two-seater cars in Table 1.6 (page 70), arranged in increasing order, are

13 15 16 16 17 19 20 22 23 23 | 23 24 25 25 26 28 28 28 29 32

The median is midway between the center pair of observations. We have marked its position in the list by |. The first quartile is the median of the 10 observations to the left of the position of the median. Check that its value is $Q_1 = 18$. Similarly, the third quartile is the median of the 10 observations to the right of the |. Check that $Q_3 = 27$.

When there is an odd number of observations, the median is the unique center observation, and the rule for finding the quartiles excludes this center value. The highway mileages of the 13 minicompact cars in Table 1.6 are (in order)

19 22 23 23 23 26 26 27 28 29 29 31 32

The median is the bold 26. The first quartile is the median of the 6 observations falling to the left of this point in the list, $Q_1 = 23$. Similarly, $Q_3 = 29$.

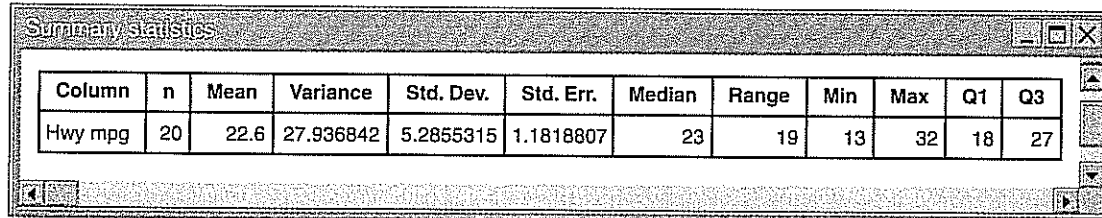
We find other percentiles more informally if we are working without software. For example, we take the 90th percentile of the 13 minicompact mileages to be the 12th in the ordered list, because $0.90 \times 13 = 11.7$, which we round to 12. The 90th percentile is therefore 31 mpg.

Example 1.14 *CrunchIt! and Minitab* Numerical summaries with computer software

Statistical software often provides several numerical measures in response to a single command. Figure 1.17 displays such output from the CrunchIt! and Minitab software for the highway mileages of two-seater cars (without the Honda Insight). Both tell us that there are

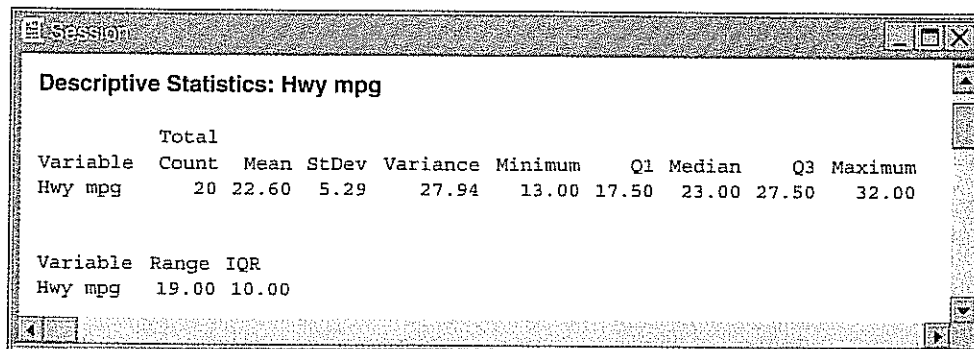
Figure 1.17 Numerical descriptions of the highway gas mileage of two-seater cars from CrunchIt! and Minitab software.

CrunchIt!



Column	n	Mean	Variance	Std. Dev.	Std. Err.	Median	Range	Min	Max	Q1	Q3
Hwy mpg	20	22.6	27.936842	5.2855315	1.1818807	23	19	13	32	18	27

Minitab



Variable	Count	Mean	StDev	Variance	Minimum	Q1	Median	Q3	Maximum
Hwy mpg	20	22.60	5.29	27.94	13.00	17.50	23.00	27.50	32.00

Variable	Range	IQR
Hwy mpg	19.00	10.00

20 observations and give the mean, median, quartiles, and smallest and largest data values. Both also give other measures, some of which we will meet soon. CrunchIt! is basic online software that offers no choice of output. Minitab allows you to choose the descriptive measures you want from a long list.

The quartiles from CrunchIt! agree with our values from Example 1.13. But Minitab's quartiles are a bit different. For example, our rule for hand calculation gives first quartile $Q_1 = 18$. Minitab's value is $Q_1 = 17.5$. There are several rules for calculating quartiles, which often give slightly different values. The differences are always small. For describing data, just report the values that your software gives.



The Five-Number Summary and Boxplots

In Section 1.1, we used the smallest and largest observations to indicate the spread of a distribution. These single observations tell us little about the distribution as a whole, but they give information about the tails of the distribution that is missing if we know only Q_1 , M , and Q_3 . To get a quick summary of both center and spread, combine all five numbers.

The Five-Number Summary

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

Minimum	Q_1	M	Q_3	Maximum
---------	-------	-----	-------	---------

These five numbers offer a reasonably complete description of center and spread. The five-number summaries for highway gas mileages are

13 18 23 27 32

for two-seaters and

19 23 26 29 32

for minicompacts. The median describes the center of the distribution; the quartiles show the spread of the center half of the data; the minimum and maximum show the full spread of the data. The five-number summary leads to another visual representation of a distribution, the **boxplot**. Figure 1.18 shows boxplots for both city and highway gas mileages for our two groups of cars.

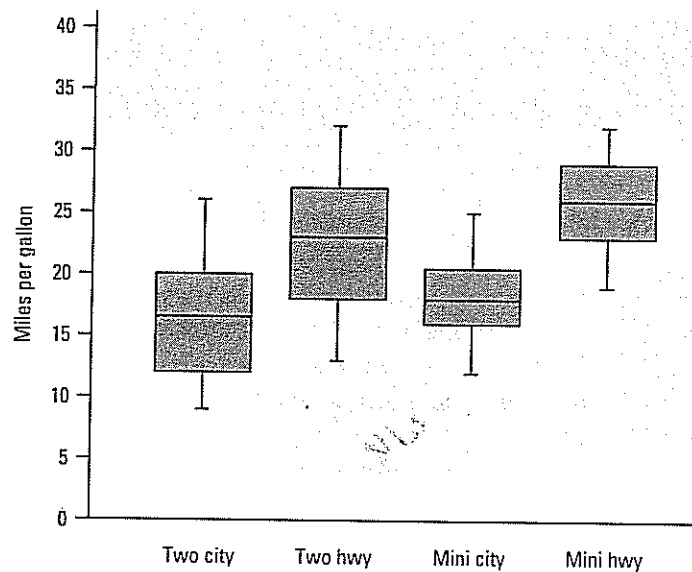
Boxplot

A boxplot is a graph of the five-number summary.

- A central box spans the quartiles Q_1 and Q_3 .
- A line in the box marks the median M .
- Lines extend from the box out to the smallest and largest observations.

Figure 1.18

Boxplots of the highway and city gas mileages for cars classified as two-seaters and as minicompacts by the Environmental Protection Agency.



Because boxplots show less detail than histograms or stemplots, they are best used for side-by-side comparison of more than one distribution, as in Figure 1.18. When you look at a boxplot, first locate the median, which marks the center of the distribution. Then look at the spread. The quartiles show the spread of the middle half of the data, and the extremes (the smallest and largest observations) show the spread of the entire data set. We see at once that city mileages are lower than highway mileages. The minicompact cars have slightly higher median gas mileages than the two-seaters, and their mileages are markedly less variable. In particular, the low gas mileages of the Ferraris and Lamborghinis in the two-seater group pull the group minimum down.

The $1.5 \times IQR$ Rule for Suspected Outliers

Look again at the stemplot of the distribution of tuition and fees for the 37 Virginia four-year colleges. Visualize the stemplot in Figure 1.4 (page 45) without the two-year college costs in the top row. You can check that the five-number summary is

9420 14,286 16,870 21,707.50 34,850

There is a clear outlier, the University of Richmond's \$34,850. How shall we describe the spread of this distribution? The range is a bit misleading because of the high outlier. The distance between the quartiles (the range of the center half of the data) is a more resistant measure of spread. This distance is called the *interquartile range*.

The Interquartile Range (IQR)

The interquartile range (*IQR*) is the distance between the first and third quartiles,

$$IQR = Q_3 - Q_1$$



For our data on Virginia tuition and fees, $IQR = 21,707.5 - 14,286 = 7421.5$. The quartiles and the *IQR* are not affected by changes in either tail of the distribution. They are therefore resistant, because changes in a few data points have no further effect once these points move outside the quartiles. However, **no single numerical measure of spread, such as *IQR*, is very useful for describing skewed distributions.** The two sides of a skewed distribution have different spreads, so one number can't summarize them. We can often detect skewness from the five-number summary by comparing how far the first quartile and the minimum are from the median (left tail) with how far the third quartile and the maximum are from the median (right tail). The interquartile range is mainly used as the basis for a rule of thumb for identifying suspected outliers.

The $1.5 \times IQR$ Rule for Outliers

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

Example 1.15 Virginia tuition and fees data $1.5 \times IQR$ rule

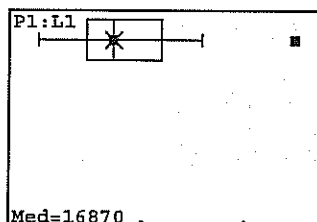
For the Virginia four-year college data in Table 1.2 (page 45),

$$1.5 \times IQR = 1.5 \times 7421.5 = 11,132.25$$

Any values below $Q_1 - 1.5 IQR = 14,286 - 11,132.25 = 3,153.75$ or above $Q_3 + 1.5 IQR = 21,707.5 + 11,132.25 = 32,839.75$ are flagged as outliers. There are no low outliers, but the University of Richmond observation, 34,850, is a high outlier.

modified boxplot

Statistical software often uses the $1.5 \times IQR$ rule. Boxplots drawn by software are often **modified boxplots** that plot suspected outliers individually. Figure 1.19 is a modified boxplot of the Virginia four-year college tuition and fees data. The lines extend out from the central box only to the smallest and largest observations that are not flagged by the $1.5 \times IQR$ rule. The largest observation is plotted as an individual point.

Figure 1.19*Boxplot of the Virginia four-year college tuition and fees data.*

The stemplot in Figure 1.4 (page 45) and the modified boxplot in Figure 1.19 tell us much more about the distribution of Virginia college tuition and fees than the five-number summary or other numerical measures. The routine methods of statistics compute numerical measures and draw conclusions based on their values. These methods are very useful, and we will study them carefully in later chapters. But they should not be applied blindly, by feeding data to a computer program, because statistical measures and methods based on them are generally meaningful only for distributions of sufficiently regular shape. This principle will become clearer as we progress, but it is good to be aware at the beginning that quickly resorting to fancy calculations is the mark of a statistical amateur. Look, think, and choose your calculations selectively.

Technology Toolbox



Calculator boxplots and numerical summaries

The TI-83/84 and TI-89 can plot up to three boxplots in the same viewing window. Both calculators can also calculate the mean, median, quartiles, and other one-variable statistics for data stored in lists. In this example, we compare Barry Bonds to Babe Ruth, the “Sultan of Swat.” Here are the numbers of home runs hit by Ruth in each of his seasons as a New York Yankee (1920 to 1934):

54 59 35 41 46 25 47 60 54 46 49 46 41 34 22

Bonds’s home runs are shown in Exercise 1.32 (page 75).

1. Enter Bonds’s home run data in L_1 /list1 and Ruth’s in L_2 /list2.
2. Set up two statistics plots: Plot 1 to show a modified boxplot of Bonds’s data and Plot 2 to show a modified boxplot of Ruth’s data.

TI-83/84

TI-89

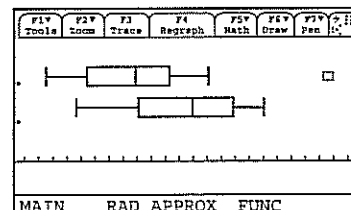
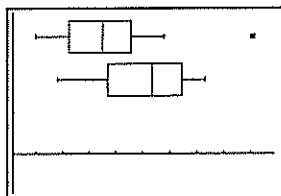
Technology Toolbox



Calculator boxplots and numerical summaries (continued)

3. Use the calculator's zoom feature to display the side-by-side boxplots.

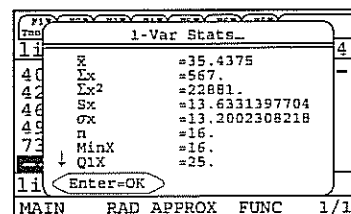
- Press **ZOOM** and select 9:ZoomStat.
- Press **F5** (ZoomData).



4. Calculate numerical summaries for each set of data.

- Press **STAT** **▶** (CALC) and select 1:1-Var Stats
- Press **ENTER**. Now press **2nd** **1** (L1) and **ENTER**.
- Press **F4** (Calc) and choose 1:1-Var Stats.
- Type list1 in the list box. Press **ENTER**.

```
1-Var Stats
x̄=35.4375
Σx=567
Σx²=22881
Sx=13.63313977
σx=13.20023082
↓n=16
```



5. Notice the down-arrow on the left side of the display. Press **▼** to see Bonds's other statistics. Repeat the process to find the Babe's numerical summaries.

Exercises

1.33 SSHA scores, II Here are the scores on the Survey of Study Habits and Attitudes (SSHA) for 18 first-year college women:

154 109 137 115 152 140 154 178 101 103 126 126 137 165 165 129 200 148

and for 20 first-year college men:

108 140 114 91 180 115 126 92 169 146 109 132 75 88 113 151 70 115 187 104

(a) Make side-by-side boxplots to compare the distributions.

- (b) Compute numerical summaries for these two distributions.
- (c) Write a paragraph comparing the SSHA scores for men and women.

1.34 How old are presidents? Return to the data on presidential ages in Table 1.4 (page 57). In Exercise 1.11, you were asked to construct a histogram of the age data.

(a) From the shape of the histogram, do you expect the mean to be much less than the median, about the same as the median, or much greater than the median? Explain.

(b) Find the five-number summary and verify your expectation from (a).

(c) What is the range of the middle half of the ages of new presidents?

(d) Construct by hand a (modified) boxplot of the ages of new presidents.

(e) On your calculator, define Plot 1 to be a histogram using the list named PREZ that you created in the Technology Toolbox on page 59. Define Plot 2 to be a (modified) boxplot also using the list PREZ. Use the calculator's zoom command to generate a graph. To remove the overlap, adjust your viewing window so that Ymin = -6 and Ymax = 22. Then graph. Use TRACE to inspect values. Press the up and down cursor keys to toggle between plots. Is there an outlier? If so, who was it?

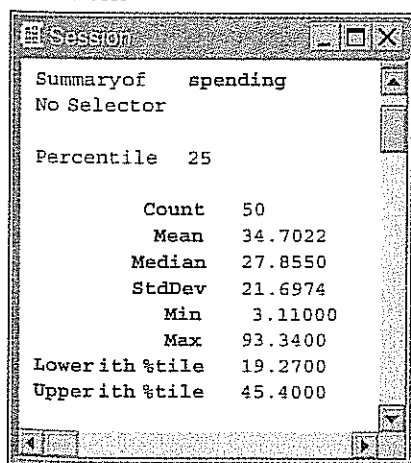
1.35 Is the interquartile range a resistant measure of spread? Give an example of a small data set that supports your answer.

1.36 Shopping spree, IV Figure 1.20 displays computer output for the data on amount spent by grocery shoppers in Exercise 1.6 (page 48).

Figure 1.20

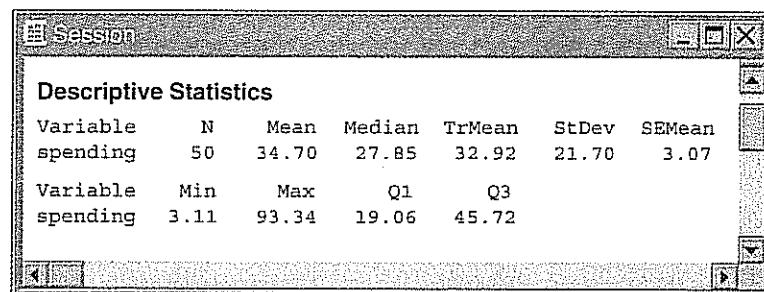
Numerical descriptions of the unrounded shopping data from the DataDesk and Minitab software, for Exercise 1.36.

DataDesk



Summary of spending	
No Selector	
Percentile 25	
Count	50
Mean	34.7022
Median	27.8550
StdDev	21.6974
Min	3.11000
Max	93.3400
Lower ith %tile	19.2700
Upper ith %tile	45.4000

Minitab



Descriptive Statistics						
Variable	N	Mean	Median	TrMean	StDev	SEMean
spending	50	34.70	27.85	32.92	21.70	3.07
Variable	Min	Max	Q1	Q3		
spending	3.11	93.34	19.06	45.72		

- (a) Find the total amount spent by the shoppers.
- (b) Make a boxplot from the computer output. Did you check for outliers?

1.37 Bonds's home runs

- (a) Find the quartiles Q_1 and Q_3 for Barry Bonds's home run data. Refer to Exercise 1.32 (page 75).
- (b) Use the $1.5 \times IQR$ rule for identifying outliers to see if Bonds's 73 home runs in 2001 is an outlier.

1.38 Senior citizens, III The stemplot you made for Exercise 1.23 (page 68) displays the distribution of the percents of residents aged 65 and over in the 50 states. Stemplots help you find the five-number summary because they arrange the observations in increasing order.

- (a) Give the five-number summary of this distribution.
- (b) Does the $1.5 \times IQR$ rule identify Alaska and Florida as suspected outliers? Does it also flag any other states?

Measuring Spread: The Standard Deviation

The five-number summary is not the most common numerical description of a distribution. That distinction belongs to the combination of the mean to measure center and the **standard deviation** to measure spread. The standard deviation measures spread by looking at how far the observations are from their mean.

The Variance s^2 and Standard Deviation s

The variance s^2 of a set of observations is the average of the squares of the deviations of the observations from their mean. In symbols, the variance of n observations x_1, x_2, \dots, x_n is

$$s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n - 1}$$

or, more compactly,

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

The standard deviation s is the square root of the variance s^2 :

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

The idea behind the variance and the standard deviation as measures of spread is as follows. The deviations $x_i - \bar{x}$ display the spread of the values x_i about their mean \bar{x} . Some of these deviations will be positive and some negative because some of the observations fall on each side of the mean. In fact, *the sum of the deviations of the observations from their mean will always be zero*. Squaring the deviations makes them all positive, so that observations far from the mean in either direction have large positive squared deviations. The variance is the average squared deviation. Therefore, s^2 and s will be large if the observations are widely spread about their mean, and small if the observations are all close to the mean.

Example 1.16 Metabolic rate

Standard deviation

A person's metabolic rate is the rate at which the body consumes energy. Metabolic rate is important in studies of weight gain, dieting, and exercise. Here are the metabolic rates of 7 men who took part in a study of dieting. (The units are calories per 24 hours. These are the same calories used to describe the energy content of foods.)

1792 1666 1362 1614 1460 1867 1439

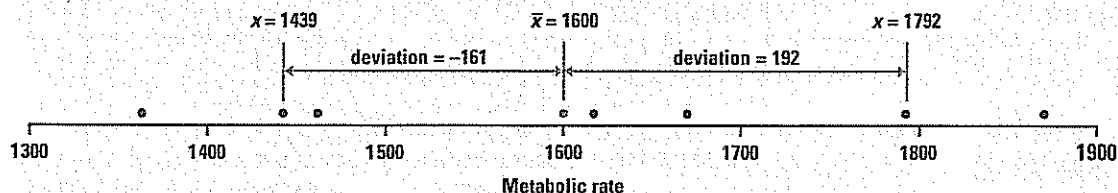
Enter these data into your calculator or software and verify that

$$\bar{x} = 1600 \text{ calories} \quad s = 189.24 \text{ calories}$$

Figure 1.21 plots these data as dots on the calorie scale, with their mean marked by the blue dot. The arrows mark two of the deviations from the mean. If you were calculating s by hand, you would find the first deviation as

$$x_1 - \bar{x} = 1792 - 1600 = 192$$

Figure 1.21 Metabolic rates for seven men, with the mean (blue dot) and the deviations of two observations from the mean indicated.



Exercise 1.41 (page 89) asks you to calculate the seven deviations, square them, and find s^2 and s directly from the deviations. Working one or two short examples by hand helps you understand how the standard deviation is obtained. In practice, you will use either software or a calculator that will find s from keyed-in data. The two software outputs in Figure 1.17 (page 77) both give the variance and standard deviation for the highway mileage data.

The idea of the variance is straightforward: it is the average of the squares of the deviations of the observations from their mean. The details we have just presented, however, raise some questions.

Why do we square the deviations? Why not just average the distances of the observations from their mean? There are two reasons, neither of them obvious. First, the sum of the squared deviations of any set of observations from their mean is the smallest such sum possible. The sum of the unsquared distances is always zero. So squared deviations point to the mean as center in a way that distances do not. Second, the standard deviation turns out to be the natural measure of spread for a particularly important class of symmetric unimodal distributions, the *Normal distributions*. We will meet the Normal distributions in the next chapter. We commented earlier that the usefulness of many statistical procedures is tied to distributions of particular shapes. This is distinctly true of the standard deviation.

Why do we emphasize the standard deviation rather than the variance? One reason is that s , not s^2 , is the natural measure of spread for Normal distributions. There is also a more general reason to prefer s to s^2 . Because the variance involves squaring the deviations, it does not have the same unit of measurement as the original observations. The variance of the metabolic rates, for example, is measured in squared calories. Taking the square root remedies this. The standard deviation s measures spread about the mean in the original scale.

degrees of freedom

Why do we “average” by dividing by $n - 1$ rather than n in calculating the variance? Because the sum of the deviations is always zero, the last deviation can be found once we know the other $n - 1$. So we are not averaging n unrelated numbers. Only $n - 1$ of the squared deviations can vary freely, and we average by dividing the total by $n - 1$. The number $n - 1$ is called the *degrees of freedom* of the variance or standard deviation. Many calculators offer a choice between dividing by n and dividing by $n - 1$, so be sure to use $n - 1$.

Properties of the Standard Deviation

Here are the basic properties of the standard deviation s as a measure of spread.

Properties of the Standard Deviation

- s measures spread about the mean and should be used only when the mean is chosen as the measure of center.
- $s = 0$ only when there is *no spread/variability*. This happens only when all observations have the same value. Otherwise, $s > 0$. As the observations become more spread out about their mean, s gets larger.
- s , like the mean \bar{x} , is not resistant. A few outliers can make s very large.



The use of squared deviations renders s even more sensitive than \bar{x} to a few extreme observations. For example, dropping the Honda Insight from our list of two-seater cars reduces the mean highway mileage from 24.7 to 22.6 mpg. It cuts the standard deviation by more than half, from 10.8 mpg with the Insight to 5.3 mpg without it. Distributions with outliers and strongly skewed distributions have large standard deviations. The number s does not give much helpful information about such distributions.

Choosing Measures of Center and Spread

How do we choose between the five-number summary and \bar{x} and s to describe the center and spread of a distribution? Because the two sides of a strongly skewed distribution have different spreads, no single number such as s describes the spread well. The five-number summary, with its two quartiles and two extremes, does a better job.

Choosing a Summary

The five-number summary is usually better than the mean and standard deviation for describing a skewed distribution or a distribution with strong outliers. Use \bar{x} and s only for reasonably symmetric distributions that are free of outliers.

Example 1.17

Investments

Choosing summaries

A central principle in the study of investments is that taking bigger risks is rewarded by higher returns, at least on the average over long periods of time. It is usual in finance to measure risk by the standard deviation of returns, on the grounds that investments whose returns vary a lot from year to year are less predictable and therefore more risky than those whose returns don't vary much. Compare, for example, the approximate mean and standard deviation of the annual percent returns on American common stocks and U.S. Treasury bills over the period from 1950 to 2003:

Investment	Mean return	Standard deviation
Common stocks	13.2%	17.6%
Treasury bills	5.0%	2.9%

Stocks are risky. They went up more than 13% per year on the average during this period, but they dropped almost 28% in the worst year. The large standard deviation reflects the fact that stocks have produced both large gains and large losses. When you buy

a Treasury bill, on the other hand, you are lending money to the government for one year. You know that the government will pay you back with interest. That is much less risky than buying stocks, so (on the average) you get a smaller return.

Are \bar{x} and s good summaries for distributions of investment returns? Figure 1.22 displays stemplots of the annual returns for both investments. (Because stock returns are so much more spread out, a back-to-back stemplot does not work well. The stems in the stock stemplot are tens of percents; the stems for Treasury bills are percents. The lowest returns are -28% for stocks and 0.9% for bills.) You see that returns on Treasury bills have a right-skewed distribution. Convention in the financial world calls for \bar{x} and s because some parts of investment theory use them. For describing this right-skewed distribution, however, the five-number summary would be more informative.

Figure 1.22

Stemplots of annual returns for stocks and Treasury bills, 1950 to 2003. (a) Stock returns, in whole percents. (b) Treasury bill returns, in percents and tenths of a percent.

Common Stocks	Treasury bills
-2 8 1	0 9
-1 9 1 1 1 1 0	1 0 2 5 5 6 6 6 8
-0 9 6 4 3	2 1 5 7 7 9
0 0 0 0 1 2 3 8 9 9	3 0 1 1 3 5 5 8 9 9
1 1 3 3 4 4 6 6 6 7 8	4 2 4 7 7 8
2 0 1 1 2 3 4 4 4 5 7 7 9 9	5 1 1 2 2 2 5 6 6 7 8 7 9
3 0 1 1 2 3 4 6 7	6 2 4 5 6 9
4 5	7 2 7 8
5 0	8 0 4 8
	9 8
	10 4 5
	11 3
	12
	13
	14 7

(a)

(b)



Remember that a graph gives the best overall picture of a distribution. Numerical measures of center and spread report specific facts about a distribution, but they do not describe its entire shape. Numerical summaries do not disclose the presence of multiple modes or gaps, for example. Always plot your data.

Exercises

1.39 Phosphate levels The level of various substances in the blood influences our health. Here are measurements of the level of phosphate in the blood of a patient, in milligrams of phosphate per deciliter of blood, made on 6 consecutive visits to a clinic:

5.6 5.2 4.6 4.9 5.7 6.4

A graph of only 6 observations gives little information, so we proceed to compute the mean and standard deviation.

(a) Find the mean from its definition. That is, find the sum of the 6 observations and divide by 6.

(b) Find the standard deviation from its definition. That is, find the deviations of each observation from the mean, square the deviations, then obtain the variance and the standard deviation. Example 1.16 shows the method.

(c) Now enter the data into your calculator to obtain \bar{x} and s . Do the results agree with your hand calculations? Can you find a way to compute the standard deviation without using one-variable statistics?

1.40 Choosing measures of center and spread Which measure of center and spread should be used for the following data sets? In each case, write a sentence or two to explain your reasoning.

(a) The Treasury bill returns in Figure 1.22(b) (page 88).

(b) The 60 IQ scores of fifth-grade students in Example 1.6 (page 49).

(c) The 44 DRP test scores in Exercise 1.5 (page 48).

1.41 Metabolic rates Calculate the mean and standard deviation of the metabolic rates in Example 1.16 (page 85), showing each step in detail. First find the mean \bar{x} by summing the 7 observations and dividing by 7. Then find each of the deviations $x_i - \bar{x}$ and their squares. Check that the deviations have sum 0. Calculate the variance as an average of the squared deviations (remember to divide by $n - 1$). Finally, obtain s as the square root of the variance.

1.42 Median and mean Create a set of 5 positive numbers (repeats allowed) that have median 10 and mean 7. What thought process did you use to create your numbers?

1.43 Contest This is a standard deviation contest. You must choose four numbers from the whole numbers 0 to 10, with repeats allowed.

(a) Choose four numbers that have the smallest possible standard deviation.

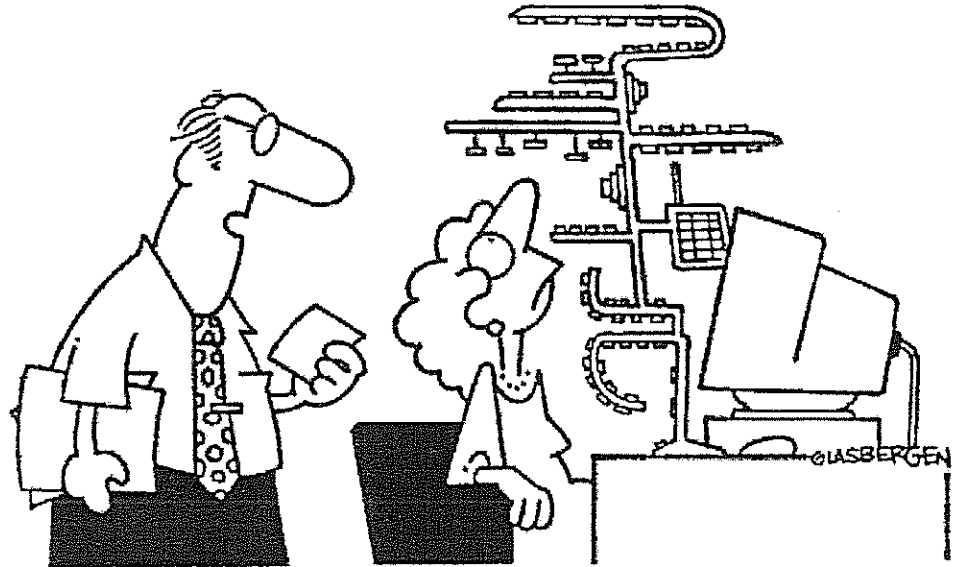
(b) Choose four numbers that have the largest possible standard deviation.

(c) Is more than one choice possible in either (a) or (b)? Explain.

1.44 Sum of deviations is zero Use the definition of the mean \bar{x} to show that the sum of the deviations $x_i - \bar{x}$ of the observations from their mean is always zero. This is one reason why the variance and standard deviation use squared deviations.



Statistics in the courtroom In 1994 Digital Equipment Corporation (DEC) was sued by three individuals who claimed that DEC's computer keyboard caused repetitive stress injuries. Awards for economic loss were fairly easy to set, but deciding awards for pain and suffering was much more difficult. On appeal, Circuit Court Judge Jack Weinstein described ways to find a comparison group of similar cases. Then for the jury award to be considered reasonable, he ruled that it should not be more than two standard deviations away from the mean award of the comparison group. Any award outside this interval would be adjusted to be two standard deviations away from the mean.



"It's the new keyboard for the statistics lab. Once you learn how to use it, it will make computation of the standard deviation easier."

Changing the Unit of Measurement

The same variable can be recorded in different units of measurement. Americans commonly record distances in miles and temperatures in degrees Fahrenheit, while the rest of the world measures distances in kilometers and temperatures in degrees Celsius. Fortunately, it is easy to convert from one unit of measurement to another. This is true because a change in the measurement unit is a **linear transformation** of the measurements.

Linear Transformation

A linear transformation changes the original variable x into the new variable x_{new} given by an equation of the form

$$x_{\text{new}} = a + bx$$

Adding the constant a shifts all values of x upward or downward by the same amount. Multiplying by the positive constant b changes the size of the unit of measurement.

Example 1.18 *Miami Heat salaries*
 Changing units


Table 1.7 gives the approximate base salaries of the 15 members of the Miami Heat basketball team for the year 2005. You can calculate that the mean is $\bar{x} = \$3.859$ million and that the median is $M = \$1.13$ million. No wonder professional basketball players have big houses!

Table 1.7 *Year 2005 salaries for Miami Heat players*
(in millions of dollars)

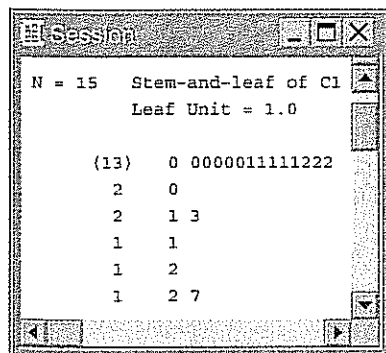
Player	Salary	Player	Salary
Shaquille O'Neal	27.70	Christian Laettner	1.10
Eddie Jones	13.46	Steve Smith	1.10
Dwyane Wade	2.83	Shandon Anderson	0.87
Damon Jones	2.50	Keyon Dooling	0.75
Michael Doleac	2.40	Zhizhi Wang	0.75
Rasual Butler	1.20	Udonis Haslem	0.62
Dorell Wright	1.15	Alonzo Mourning	0.33
Qyntel Woods	1.13		

Source: www.hoopshype.com/salaries/miami.htm.

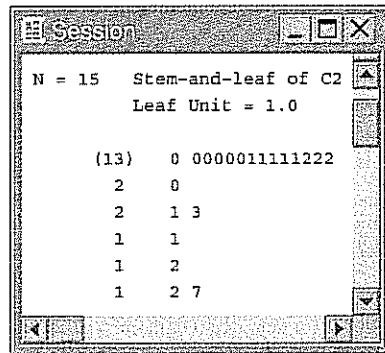
Figure 1.23(a) is a stemplot of the salaries of Miami Heat players, with millions as stems. The distribution is skewed to the right and there are two high outliers. The very high

Figure 1.23

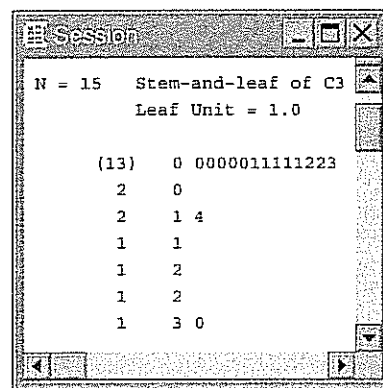
Minitab stemplots. (a) The salaries of Miami Heat players, with millions as stems. (b) When the same amount is added to every observation, the shape of the distribution remains unchanged. (c) Multiplying every observation by the same amount will either increase or decrease the spread by that amount.



(a)



(b)



(c)

salaries of Eddie Jones and Shaquille O'Neal pull up the mean. Use your calculator to check that $s = \$7.34$ million, and that the five-number summary is

\$0.33 million \$0.75 million \$1.13 million \$2.50 million \$27.70 million

1. Suppose that each member of the team receives a \$100,000 bonus for winning the NBA Championship. How will this affect the shape, center, and spread of the distribution?

Since $\$100,000 = \0.1 million, each player's salary will increase by \$0.1 million. This linear transformation can be represented by $x_{\text{new}} = 0.1 + 1x$, where x_{new} is the salary after the bonus and x is the player's base salary. Increasing each value in Table 1.7 by 0.1 will also increase the mean by 0.1. That is, $\bar{x}_{\text{new}} = \$3.96$ million. Likewise, the median salary will increase by 0.1 and become $M = \$1.23$ million.

What will happen to the spread of the distribution? The standard deviation of the Heat's salaries after the bonus is still $s = \$7.34$ million. With the bonus, the five-number summary becomes

\$0.43 million \$0.85 million \$1.23 million \$2.60 million \$27.80 million

Both before and after the salary bonus, the *IQR* for this distribution is \$1.75 million. *Adding a constant amount to each observation does not change the spread.* The shape of the distribution remains unchanged, as shown in Figure 1.23(b).

2. Suppose that, instead of receiving a \$100,000 bonus, each player is offered a 10% increase in his base salary. Alonzo Mourning, who is making a base salary of \$0.33 million, would receive an additional $(0.10)(\$0.33 \text{ million}) = \0.033 million. To obtain his new salary, we could have used the linear transformation $x_{\text{new}} = 0 + 1.10x$, since multiplying the current salary (x) by 1.10 increases it by 10%. Increasing all 15 players' salaries in the same way results in the following list of values (in millions):

\$0.363	\$0.682	\$0.825	\$0.825	\$0.957	\$ 1.210	\$ 1.210	\$ 1.243
\$1.265	\$1.320	\$2.640	\$2.750	\$3.113	\$14.806	\$30.470	

Use your calculator to check that $\bar{x}_{\text{new}} = \$4.245$ million, $s_{\text{new}} = \$8.072$ million, $M_{\text{new}} = \$1.243$ million, and the five-number summary for x_{new} is

\$0.363 million \$0.825 million \$1.243 million \$2.750 million \$30.470 million

Since $\$3.859(1.10) = \4.245 and $\$1.13(1.10) = \1.243 , you can see that both measures of center (the mean and median) have increased by 10%. This time, the spread of the distribution has increased, too. Check for yourself that the standard deviation and the *IQR* have also increased by 10%. The stemplot in Figure 1.23(c) shows that the distribution of salaries is still right-skewed.

Linear transformations do not change the shape of a distribution. If measurements on a variable x have a right-skewed distribution, any new variable x_{new} obtained by a linear transformation $x_{\text{new}} = a + bx$ (for $b > 0$) will also have a right-skewed distribution. If the distribution of x is symmetric and unimodal, the distribution of x_{new} remains symmetric and unimodal.

Although a linear transformation preserves the basic shape of a distribution, the center and spread may change. Because linear transformations of measurement scales are common, we must be aware of their effect on numerical descriptive measures of center and spread. Fortunately, the changes follow a simple pattern.

Effect of a Linear Transformation

To see the effect of a linear transformation on measures of center and spread, apply these rules:

- Multiplying each observation by a positive number b multiplies both measures of center (mean and median) and measures of spread (interquartile range and standard deviation) by b .
- Adding the same number a (either positive, zero, or negative) to each observation adds a to measures of center and to quartiles but does not change measures of spread.

The measures of spread IQR and s do not change when we add the same number a to all of the observations because adding a constant changes the location of the distribution but leaves the spread unaltered. You can find the effect of a linear transformation $x_{\text{new}} = a + bx$ by combining these rules. For example, if x has mean \bar{x} , the transformed variable x_{new} has mean $a + b\bar{x}$.

Comparing Distributions

An experiment is carried out to compare the effectiveness of a new cholesterol-reducing drug with the one that is currently prescribed by most doctors. A survey is conducted to determine whether the proportion of males who are likely to vote for a political candidate is higher than the proportion of females who are likely to vote for the candidate. Students taking AP Calculus AB and AP Statistics are curious about which exam is harder. They have information on the distribution of scores earned on each exam from the year 2005. In each of these situations, we are interested in comparing distributions. This section presents some of the more common methods for making statistical comparisons. We also introduce the Data Analysis Toolbox. When you are investigating a statistical problem involving one or more sets of data, use the Data Analysis Toolbox to organize your thinking.

Data Analysis Toolbox

To answer a statistical question of interest involving one or more data sets, proceed as follows.

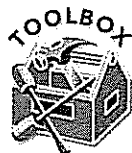
- **Data** Organize and examine the data. Answer the *key questions* from the Preliminary Chapter:
 1. Who are the individuals described by the data?
 2. What are the variables? In what units is each variable recorded?

Data Analysis Toolbox*(continued)*

3. Why were the data gathered?
 4. When, where, how, and by whom were the data produced?
- **Graphs** Construct appropriate graphical displays.
 - **Numerical summaries** Calculate relevant summary statistics.
 - **Interpretation** Discuss what the data, graphs, and numerical summaries tell you in the context of the problem. Answer the question!

Example 1.19

The nation's report card
Comparing categorical variables



The National Assessment of Educational Progress (NAEP), also known as "the Nation's Report Card," is a nationally representative and continuing assessment of what Americans know and can do in reading and math. NAEP results are based on samples of students. Table 1.8 shows the NAEP comparisons between Virginia students and the nation as a whole.

Table 1.8

How Virginia students performed on national reading and math tests.

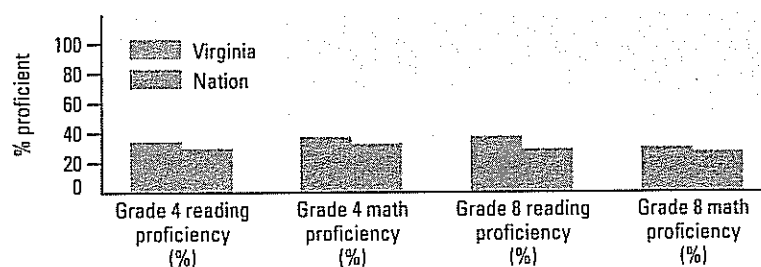
	Virginia	Nation
Grade 4 reading proficiency (%)	35	30
Grade 4 math proficiency (%)	36	31
Grade 8 reading proficiency (%)	36	30
Grade 8 math proficiency (%)	31	27

Source: Data from the Web site www.nces.ed.gov/nationsreportcard/.

- **Data**
1. **Who?** The individuals are students in grades 4 and 8.
 2. **What?** The two variables for Virginia and the nation are the percent of individuals classified as proficient on the math test and the percent classified as proficient on the reading test.
 3. **Why?** The National Center for Education Statistics is charged with periodically testing students in mathematics and reading. They are also charged with preparing reports for the public that enable comparisons of an individual's performance with that of students across the nation and in the state.
 4. **When, where, how, and by whom?** The data were collected in 2003 by the National Center for Education Statistics, based in Washington, D.C.

- **Graphs** Figure 1.24 shows how Virginia's grade 4 and grade 8 students compare with the nation as a whole in reading and math proficiency.

Figure 1.24 Bar chart comparing Virginia and the nation in reading and math proficiency.



- **Numerical summaries** are the percents of students who test proficient. See Table 1.8 on the previous page.
- **Interpretation** Table 1.8 and Figure 1.24 show that higher percents of students were proficient in math and reading in Virginia than in the nation as a whole.

Example 1.20

Swiss doctors

Comparing quantitative variables: Data Analysis Toolbox



Do male doctors perform more cesarean sections (C-sections) than female doctors? A study in Switzerland examined the number of cesarean sections (surgical deliveries of babies) performed in a year by samples of male and female doctors.

- **Data** Here are the data for 15 male doctors, arranged from lowest to highest:

20	25	25	27	28	31	33	34
36	37	44	50	59	85	86	

The study also looked at 10 female doctors. The numbers of cesarean sections performed by these doctors (arranged in order) were

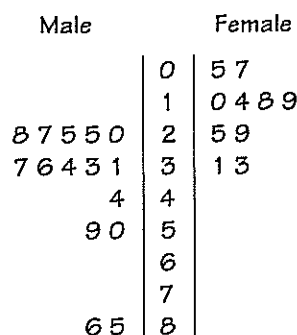
5 7 10 14 18 19 25 29 31 33

1. **Who?** The individuals are Swiss doctors.
2. **What?** The variable is the number of cesarean sections performed in a year.
3. **Why?** Researchers wanted to compare the number of cesarean sections performed by male and female doctors in Switzerland.
4. **When, where, how, and by whom?** The unpublished results of this study were provided to one of the co-authors in a private communication in the 1980s. The identities of the researchers are unknown.

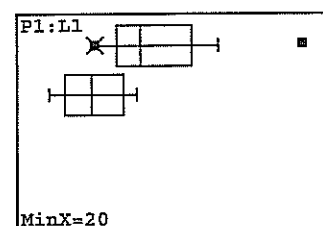
- **Graphs** We can compare the number of cesarean sections performed by male and female doctors using a back-to-back stemplot and a side-by-side boxplot. Figure 1.25 shows the completed graphs. In the stemplot, the stems are listed in the middle and leaves are placed on the left for male doctors and on the right for female doctors. It is usual to have the leaves increase in value as they move away from the stem.

Figure 1.25

(a) Back-to-back stemplot of the number of cesarean sections performed by male and female Swiss doctors. (b) Side-by-side boxplots of cesarean section data.



(a)



(b)

- **Numerical summaries** Here are summary statistics for the two distributions:

	\bar{x}	s	Min.	Q_1	M	Q_3	Max.	IQR
Male doctors	41.333	20.607	20	27	34	50	86	23
Female doctors	19.1	10.126	5	10	18.5	29	33	19

- **Interpretation** The distribution of the number of cesarean sections performed by female doctors is roughly symmetric. For the male doctors, the distribution is skewed to the right. (*Shape*)

Two male physicians performed an unusually high number of cesarean sections, 85 and 86. You can use the $1.5 \times IQR$ rule to confirm that these two values are outliers, and that there are no other outliers in either data set. (*Outliers*)

More than half of the female doctors in the study performed fewer than 20 cesarean sections in a year; 20 was the minimum number of cesarean sections performed by male doctors. The mean and median numbers of cesarean sections performed are higher for the male doctors. (*Center*)

Both the standard deviation and the IQR for the male doctors are much larger than the corresponding statistics for the female doctors. So there is much greater variability in the number of cesarean sections performed by male doctors. (*Spread*)

Due to the outliers in the male doctor data and the lack of symmetry of their distribution of cesareans, we should use the resistant medians and IQR s in our numerical comparisons. In Switzerland, it does seem that male doctors generally perform more cesarean sections each year (median = 34) than do female doctors (median = 18.5). In addition, male Swiss doctors are more variable in the number of cesarean sections performed each year ($IQR = 23$) than female Swiss doctors ($IQR = 19$). We may want to do more research on why this apparent discrepancy exists.

Exercises

1.45 Raising teachers' pay, I A school system employs teachers at salaries between \$30,000 and \$60,000. The teachers' union and the school board are negotiating the form of next year's increase in the salary schedule. Suppose that every teacher is given a flat \$1000 raise.

- (a) How much will the mean salary increase? The median salary?
- (b) Will a flat \$1000 raise increase the spread as measured by the distance between the quartiles?
- (c) Will a flat \$1000 raise increase the spread as measured by the standard deviation of the salaries?

1.46 Raising teachers' pay, II Suppose that the teachers in the previous exercise each receive a 5% raise. The amount of the raise will vary from \$1500 to \$3000, depending on present salary. Will a 5% across-the-board raise increase the spread of the distribution as measured by the distance between the quartiles? Do you think it will increase the standard deviation?

1.47 Which AP exam is easier: Calculus AB or Statistics? The table below gives the distribution of grades earned by students taking the AP Calculus AB and AP Statistics exams in 2005:

	5	4	3	2	1
Calculus AB	20.7%	19.5%	17.7%	16.7%	25.2%
Statistics	12.6%	22.8%	25.3%	19.2%	20.1%

- (a) Make a graphical display to compare the exam grades for Calculus AB and Statistics.
- (b) Write a few sentences comparing the two distributions of exam grades. Do you now know which exam is easier? Why or why not?

1.48 Get your hot dogs here! Face it. "A hot dog isn't a carrot stick." So said *Consumer Reports*, commenting on the low nutritional quality of the all-American frank. Table 1.9 on the next page shows the magazine's laboratory test results for calories and milligrams of sodium (mostly due to salt) in a number of major brands of hot dogs. There are three types: beef, "meat" (mainly pork and beef, but government regulations allow up to 15% poultry meat), and poultry. Because people concerned about their health may prefer low-calorie, low-sodium hot dogs, we ask: "Are there any systematic differences among the three types of hot dogs in these two variables?" Use side-by-side boxplots and numerical summaries to help you answer this question. Write a paragraph explaining your findings. Use the Data Analysis Toolbox (page 93) as a guide.

1.49 Who makes more? A manufacturing company is reviewing the salaries of its full-time employees below the executive level at a large plant. The clerical staff is almost entirely female, while a majority of the production workers and technical staff are male. As a result, the distributions of salaries for male and female employees may be quite different. Table 1.10 on the next page gives the frequencies and relative frequencies for women and men.

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

Table 1.9 *Calories and sodium (milligrams) in three types of hot dogs*

Beef Hot Dogs		Meat Hot Dogs		Poultry Hot Dogs	
Calories	Sodium	Calories	Sodium	Calories	Sodium
186	495	173	458	129	430
181	477	191	506	132	375
176	425	182	473	102	396
149	322	190	545	106	383
184	482	172	496	94	387
190	587	147	360	102	542
158	370	146	387	87	359
139	322	139	386	99	357
175	479	175	507	170	528
148	375	136	393	113	513
152	330	179	405	135	426
111	300	153	372	142	513
141	386	107	144	86	358
153	401	195	511	143	581
190	645	135	405	152	588
157	440	140	428	146	522
131	317	138	339	144	545
149	319				
135	298				
132	253				

Source: *Consumer Reports*, June 1986, pp. 366–367.**Table 1.10** *Salary distributions of female and male workers in a large factory*

Salary (\$1000)	Women		Men	
	Number	%	Number	%
10–15	89	11.8	26	1.1
15–20	192	25.4	221	9.0
20–25	236	31.2	677	27.6
25–30	111	14.7	823	33.6
30–35	86	11.4	365	14.9
35–40	25	3.3	182	7.4
40–45	11	1.5	91	3.7
45–50	3	0.4	33	1.4
50–55	2	0.3	19	0.8
55–60	0	0.0	11	0.4
60–65	0	0.0	0	0.0
65–70	1	0.1	3	0.1
Total	756	100.1	2451	100.0

- (b) Describe the shapes of the salary distributions and the chief differences between them.
- (c) Explain why the total for women is greater than 100%.

1.50 Linear transformations In each of the following settings, give the values of a and b for the linear transformation $x_{\text{new}} = a + bx$ that expresses the change in measurement units. Then explain how the transformation will affect the mean, the *IQR*, the median, and the standard deviation of the original distribution.

- (a) You collect data on the power of car engines, measured in horsepower. Your teacher requires you to convert the power to watts. One horsepower is 746 watts.
- (b) You measure the temperature (in degrees Fahrenheit) of your school's swimming pool at 20 different locations within the pool. Your swim team coach wants the summary statistics in degrees Celsius ($^{\circ}\text{F} = (9/5)^{\circ}\text{C} + 32$).
- (c) Mrs. Swaynos has given a very difficult statistics test and is thinking about "curving" the grades. She decides to add 10 points to each student's score.

Section 1.2 Summary

A numerical summary of a distribution should report its **center** and its **spread**, or **variability**.

The mean \bar{x} and the median M describe the center of a distribution in different ways. The mean is the average of the observations, and the median is the midpoint of the values.

When you use the median to indicate the center of a distribution, describe its spread by giving the **quartiles**. The **first quartile** Q_1 has about one-fourth of the observations below it, and the **third quartile** Q_3 has about three-fourths of the observations below it. An extreme observation is an **outlier** if it is smaller than $Q_1 - (1.5 \times IQR)$ or larger than $Q_3 + (1.5 \times IQR)$.

The **five-number summary** consists of the median, the quartiles, and the high and low extremes and provides a quick overall description of a distribution. The median describes the center, and the quartiles and extremes show the spread.

Boxplots based on the five-number summary are useful for comparing two or more distributions. The box spans the quartiles and shows the spread of the central half of the distribution. The median is marked within the box. Lines extend from the box to the smallest and the largest observations that are not outliers. Outliers are plotted as isolated points.

The **variance** s^2 and especially its square root, the **standard deviation** s , are common measures of spread about the mean as center. The standard deviation s is zero when there is no spread and gets larger as the spread increases.

The median is a **resistant** measure of center because it is relatively unaffected by extreme observations. The mean is nonresistant. Among measures of spread, the quartiles are resistant, but the standard deviation is not.

The mean and standard deviation are strongly influenced by outliers or skewness in a distribution. They are good descriptions for symmetric distributions and are most useful for the Normal distributions, which will be introduced in the next chapter.

The median and quartiles are not affected by outliers, and the two quartiles and two extremes describe the two sides of a distribution separately. The five-number summary is the preferred numerical summary for skewed distributions.

Linear transformations are quite useful for changing units of measurement. Linear transformations have the form $x_{\text{new}} = a + bx$. When you add a constant a to all the values in a data set, the mean and median increase by a . Measures of spread do not change. When you multiply all the values in a data set by a constant b , the mean, median, IQR, and standard deviation are multiplied by b .

Back-to-back stemplots and **side-by-side boxplots** are useful for comparing quantitative distributions.

Section 1.2 Exercises

1.51 Do girls study more than boys? We asked the students in three AP Statistics classes how many minutes they studied on a typical weeknight. Here are the responses of random samples of 30 girls and 30 boys from the classes:

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

(a) Examine the data. Why are you not surprised that most responses are multiples of 10 minutes? We eliminated one student who claimed to study 30,000 minutes per night. Are there any other responses you consider suspicious?

(b) Make a back-to-back stemplot of these data. Report the approximate midpoints of both groups. Does it appear that girls study more than boys (or at least claim that they do)?

1.52 Educational attainment Table 1.11 on the next page shows the educational level achieved by U.S. adults aged 25 to 34 and by those aged 65 to 74. Compare the distributions of educational attainment graphically. Write a few sentences explaining what your display shows.

1.53 Logging in rain forests "Conservationists have despaired over destruction of tropical rainforest by logging, clearing, and burning." These words begin a report on a statistical study of the effects of logging in Borneo. Researchers compared forest plots that had never been logged (Group 1) with similar plots nearby that had been logged 1 year earlier (Group 2) and 8 years earlier (Group 3). All plots were 0.1 hectare in area. Here are the counts of trees for plots in each group:¹⁸

Group 1:	27	22	29	21	19	33	16	20	24	27	28	19
Group 2:	12	12	15	9	20	18	17	14	14	2	17	19
Group 3:	18	4	22	15	18	19	22	12	12			

(a) Give a complete comparison of the three distributions, using both graphs and numerical summaries.

Table 1.11 Educational attainment by U.S. adults aged 25 to 34 and 65 to 74

	Number of People (thousands)	
	Ages 25–34	Ages 65–74
Less than high school	5,063	4,546
High school graduate	11,380	6,737
Some college	7,613	2,481
Bachelor's degree	8,830	2,047
Advanced degree	2,943	1,413
Total	35,829	17,224

Source: Census Bureau, *Educational Attainment in the United States*, March 2004.

(b) To what extent has logging affected the count of trees?

(c) The researchers used an analysis based on \bar{x} and s . Explain why this is reasonably well justified.

1.54 \bar{x} and s are not enough The mean \bar{x} and standard deviation s measure center and spread but are not a complete description of a distribution. Data sets with different shapes can have the same mean and standard deviation. To demonstrate this fact, use your calculator to find \bar{x} and s for the following two small data sets. Then make a stemplot of each and comment on the shape of each distribution.

Data A:	9.14	8.14	8.74	8.77	9.26	8.10	6.13	3.10	9.13	7.26	4.74
Data B:	6.58	5.76	7.71	8.84	8.47	7.04	5.25	5.56	7.91	6.89	12.50

1.55 Sales of CDs In recent years, the Recording Industry Association of America (RIAA) has initiated legal action against individuals for illegally downloading copyrighted music from the Web. RIAA has targeted primarily college students, roughly aged 17 to 23, who access peer-to-peer (P2P) sites via fast Internet connections in their dorm rooms. The table below shows the sales from record labels, primarily CDs (87.8% of sales), from 1994 to 2003 for two different age groups.¹⁹

	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003
15–34 years	44.8	44.7	44.7	42.3	39.4	35.7	36.0	36.1	34.2	32.3
Over 35	46.6	46.5	47.3	47.9	50.4	54.5	53.8	54.5	56.0	57.9

Make a graphical display to compare sales from record labels to these two consumer groups. Write a few sentences describing what the graph tells you.

1.56 Linear transformation consequences A change of units that multiplies each unit by b , such as the change $x_{\text{new}} = 0 + 2.54x$ from inches x to centimeters x_{new} , multiplies our usual measures of spread by b . This is true of the IQR and standard deviation. What happens to the variance when we change units in this way?

1.57 Better corn Corn is an important animal food. Normal corn lacks certain amino acids, which are building blocks for protein. Plant scientists have developed new corn



varieties that have more of these amino acids. To test a new corn as an animal food, a group of 20 one-day-old male chicks was fed a ration containing the new corn. A control group of another 20 chicks was fed a ration that was identical except that it contained normal corn. Here are the weight gains (in grams) after 21 days:²⁰

Normal corn				New corn			
380	321	366	356	361	447	401	375
283	349	402	462	434	403	393	426
356	410	329	399	406	318	467	407
350	384	316	272	427	420	477	392
345	455	360	431	430	339	410	326

(a) Compute five-number summaries for the weight gains of the two groups of chicks. Then make boxplots to compare the two distributions. What do the data show about the effect of the new corn?

(b) The researchers actually reported means and standard deviations for the two groups of chicks. What are they? How much larger is the mean weight gain of chicks fed the new corn?

(c) The weights are given in grams. There are 28.35 grams in an ounce. Use the results of part (b) to compute the means and standard deviations of the weight gains measured in ounces.

1.58 Mean or median? Which measure of center, the mean or the median, should you use in each of the following situations?

(a) Middletown is considering imposing an income tax on citizens. The city government wants to know the average income of citizens so that it can estimate the total tax base.

(b) In a study of the standard of living of typical families in Middletown, a sociologist estimates the average family income in that city.

C A S E C L O S E D !

Nielsen ratings

Begin by reviewing the ratings data in the Nielsen ratings Case Study (page 37). Then answer each of the following questions in complete sentences. Be sure to communicate clearly enough for any of your classmates to understand what you are saying.

1. Construct by hand an appropriate graphical display for comparing the Nielsen ratings of the three networks. Write a few sentences describing what you see.

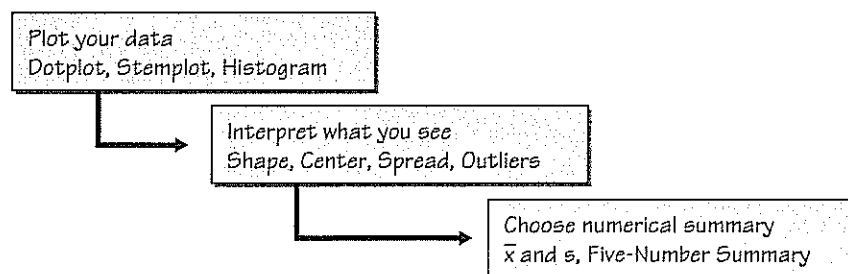
2. Calculate numerical summaries for the Nielsen ratings of the three networks. Which measures of center and spread would you choose to compare the distributions and why?
3. Determine whether there are any outliers in each of the three distributions. If there are outliers, identify them.
4. What does it mean to say that the mean percent of TV viewers watching a particular network is a nonresistant measure of center?
5. How would you rank the three networks based on your analysis?



Chapter Review

Summary

Data analysis is the art of describing data using graphs and numerical summaries. The purpose of data analysis is to describe the most important features of a set of data. This chapter introduces data analysis by presenting statistical ideas and tools for describing the distribution of a single variable. The figure below will help you organize the big ideas.



Chapter Review Exercises

1.59 Top companies Each year *Fortune* magazine lists the top 500 companies in the United States, ranked according to their total annual sales in dollars. Describe three other variables that could reasonably be used to measure the “size” of a company.

1.60 Density of the earth In 1798 the English scientist Henry Cavendish measured the density of the earth by careful work with a torsion balance. The variable recorded was the density of the earth as a multiple of the density of water. Here are Cavendish’s 29 measurements:²¹

5.50	5.61	4.88	5.07	5.26	5.55	5.36	5.29	5.58	5.65
5.57	5.53	5.62	5.29	5.44	5.34	5.79	5.10	5.27	5.39
5.42	5.47	5.63	5.34	5.46	5.30	5.75	5.68	5.85	

Present these measurements graphically in a stemplot. Discuss the shape, center, and spread of the distribution. Are there any outliers? What is your estimate of the density of the earth based on these measurements?

1.61 Hummingbirds and tropical flowers Different varieties of the tropical flower *Heliconia* are fertilized by different species of hummingbirds. Over time, the lengths of the flowers and the forms of the hummingbirds’ beaks have evolved to match each other. Here are data on the lengths in millimeters of three varieties of these flowers on the island of Dominica:²²

H. bihai

47.12	46.75	46.80	47.12	46.67	47.43	46.44	46.64
48.07	48.34	48.15	50.26	50.12	46.34	46.94	48.36

H. caribaea red

41.90	42.01	41.93	43.09	41.47	41.69	39.78	40.57
39.63	42.18	40.66	37.87	39.16	37.40	38.20	38.07
38.10	37.97	38.79	38.23	38.87	37.78	38.01	

H. caribaea yellow

36.78	37.02	36.52	36.11	36.03	35.45	38.13	37.10
35.17	36.82	36.66	35.68	36.03	34.57	34.63	

(a) Make boxplots to compare the three distributions. Report the five-number summaries along with your graphs. What are the most important differences among the three varieties of flower?

(b) Find \bar{x} and s for each variety.

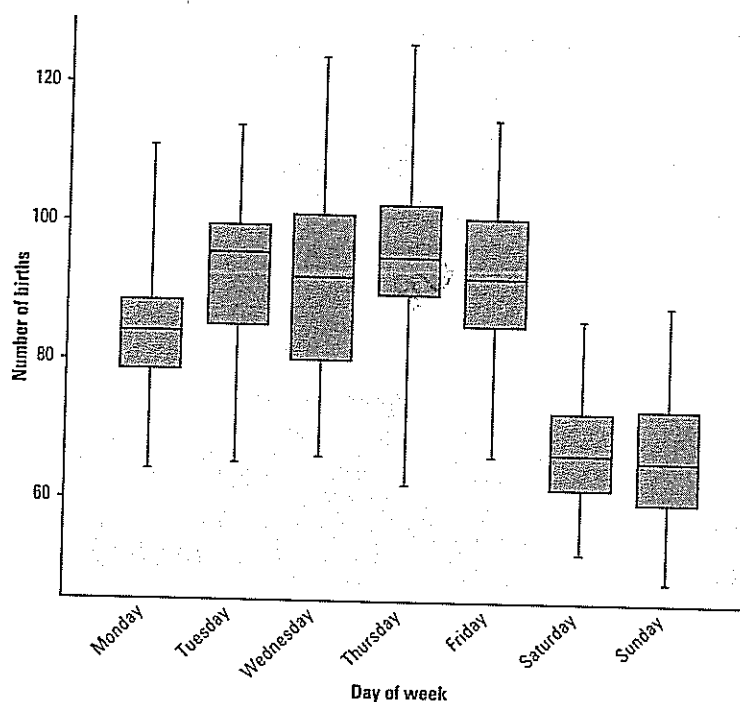
(c) Make a stemplot of each set of flower lengths. Do the distributions appear suitable for use of \bar{x} and s as summaries?

(d) Starting from the \bar{x} and s -values in millimeters, find the means and standard deviations in inches. (A millimeter is 1/1000 of a meter. A meter is 39.37 inches.)

1.62 Never on Sunday Figure 1.26 shows the distributions of number of births in Toronto, Canada, on each of the 365 days in a year, grouped by day of the week.²³ Based on these plots, give a more detailed description of how births depend on the day of the week.

Figure 1.26

Side-by-side boxplots of the distributions of numbers of births in Toronto, Canada, for each day of the week during a year, for Exercise 1.62.



1.63 Presidential elections Here are the percents of the popular vote won by the successful candidate in each of the presidential elections from 1948 to 2004:

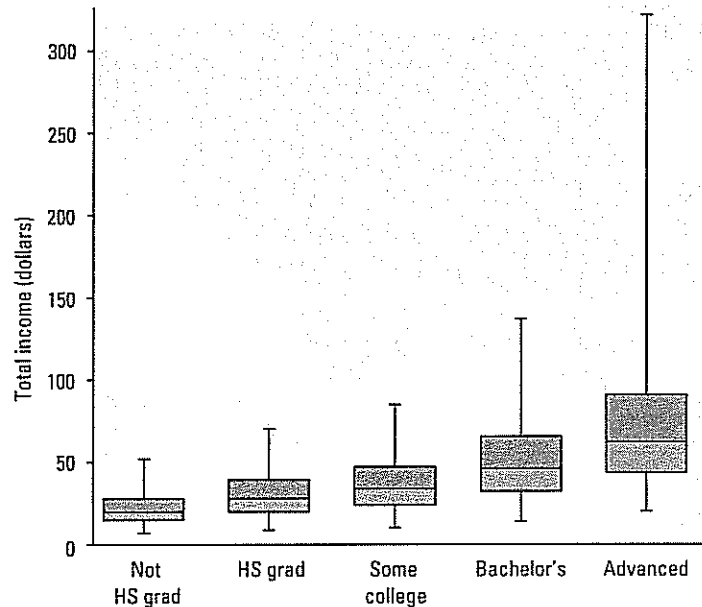
Year:	1948	1952	1956	1960	1964	1968	1972	1976	1980	1984	1988	1992	1996	2000	2004
Percent:	49.6	55.1	57.4	49.7	61.1	43.4	60.7	50.1	50.7	58.8	53.9	43.2	49.2	47.9	50.7

- Make a stemplot of the winners' percents. (Round to whole numbers and use split stems.)
- What is the median percent of the vote won by the successful candidate in presidential elections? (Work with the unrounded data.)
- Call an election a landslide if the winner's percent falls at or above the third quartile. Find the third quartile. Which elections were landslides?

1.64 Income and education level Each March, the Bureau of Labor Statistics compiles an Annual Demographic Supplement to its monthly Current Population Survey.²⁴ Data on about 71,067 individuals between the ages of 25 and 64 who were employed full-time in 2001 were collected in one of these surveys. The boxplots in Figure 1.27 compare the distributions of income for people with five levels of education. This figure is a variation

Figure 1.27

Boxplots comparing the distributions of income for employed people aged 25 to 64 years with five different levels of education, for Exercise 1.64. The lines extend from the quartiles to the 5th and 95th percentiles.



of the boxplot idea: because large data sets often contain very extreme observations, the lines extend from the central box only to the 5th and 95th percentiles. The data include 14,959 people whose highest level of education is a bachelor's degree.

- What is the position of the median in the ordered list of incomes (1 to 14,959) of people with a bachelor's degree? From the boxplot, about what is the median income?
- What is the position of the first and third quartiles in the ordered list of incomes for these people? About what are the numerical values of Q_1 and Q_3 ?
- You answered (a) and (b) from a boxplot that omits the lowest 5% and the highest 5% of incomes. Explain why leaving out these values has only a very small effect on the median and quartiles.
- About what are the positions of the 5th and 95th percentiles in the ordered list of incomes of the 14,959 people with a bachelor's degree? Incomes outside this range do not appear in the boxplot.
- About what are the numerical values of the 5th and 95th percentiles of income? (For comparison, the largest income among all 14,959 people was \$481,720. That one person made this much tells us less about the group than does the 95th percentile.)
- Write a brief description of how the distribution of income changes with the highest level of education reached. Be sure to discuss center, spread, and skewness. Give some specifics read from the graphs to back up your statements.

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

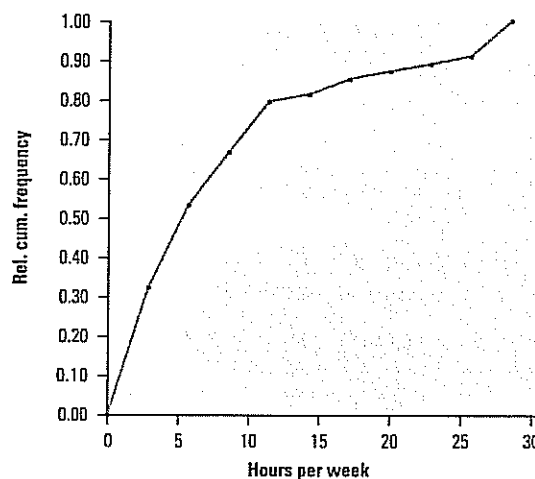
8.25	7.83	8.30	8.42	8.50	8.67	8.17	9.00	9.00	8.17	7.92
9.00	8.50	9.00	7.75	7.92	8.00	8.08	8.42	8.75	8.08	9.75
8.33	7.83	7.92	8.58	7.83	8.42	7.75	7.42	6.75	7.42	8.50
8.67	10.17	8.75	8.58	8.67	9.17	9.08	8.83	8.67		

- Make a histogram of these drive times. Is the distribution roughly symmetric, clearly skewed, or neither? Are there any clear outliers?
- Construct an ogive for Professor Moore's drive times.
- Use your ogive from (b) to estimate the center and 90th percentile of the distribution.
- Use your ogive to estimate the percentile corresponding to a drive time of 8.00 minutes.

1.66 Computer use Mrs. Causey asked her students how much time they had spent using a computer during the previous week. Figure 1.28 is an ogive of her students' responses.

Figure 1.28

Ogive of weekly computer use by Mrs. Causey's students, for Exercise 1.66.



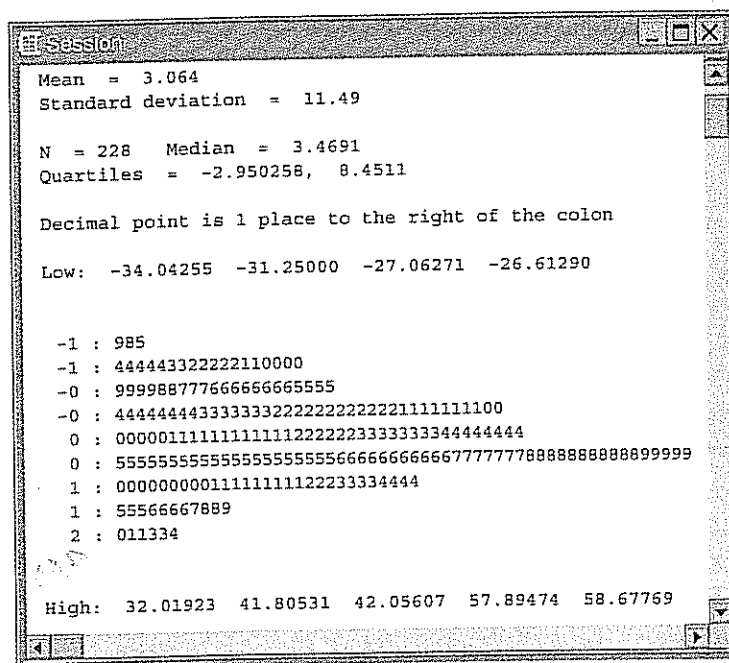
- Construct a relative frequency table based on the ogive. Then make a histogram.
- Estimate the median, Q_1 , and Q_3 from the ogive. Then make a boxplot. Are there any outliers?
- At what percentile does a student who used her computer for 10 hours last week fall?

1.67 Wal-Mart stock The rate of return on a stock is its change in price plus any dividends paid. Rate of return is usually measured in percent of the starting value. We have data on the monthly rates of return for the stock of Wal-Mart stores for the years 1973 to 1991, the

first 19 years Wal-Mart was listed on the New York Stock Exchange. There are 228 observations. Figure 1.29 displays output from statistical software that describes the distribution of these data. The stems in the stemplot are the tens digits of the percent returns. The leaves are the ones digits. The stemplot uses split stems to give a better display. The software gives high and low outliers separately from the stemplot rather than spreading out the stemplot to include them.

Figure 1.29

Output from Minitab software describing the distribution of monthly returns from Wal-Mart stock, for Exercise 1.67.



- Give the five-number summary for monthly returns on Wal-Mart stock.
- Describe in words the main features of the distribution.
- If you had \$1000 worth of Wal-Mart stock at the beginning of the best month during these 19 years, how much would your stock be worth at the end of the month? If you had \$1000 worth of stock at the beginning of the worst month, how much would your stock be worth at the end of the month?
- Find the interquartile range (IQR) for the Wal-Mart data. Are there any outliers according to the $1.5 \times IQR$ criterion? Does it appear to you that the software uses this criterion in choosing which observations to report separately as outliers?

1.68 Jury awards A study of the size of jury awards in civil cases (such as injury, product liability, and medical malpractice) in Chicago showed that the median award was about \$8000. But the mean award was about \$69,000. Explain how a difference this big between the two measures of center can occur.

1.69 Women runners Women were allowed to enter the Boston Marathon in 1972. Here are the times (in minutes, rounded to the nearest minute) for the winning woman from 1972 to 2003:

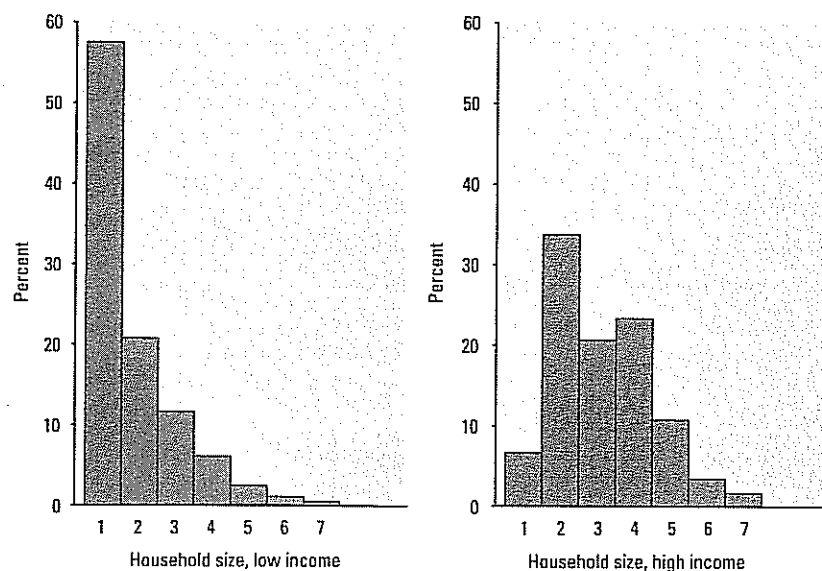
Year	Time	Year	Time	Year	Time	Year	Time
1972	190	1980	154	1988	145	1996	147
1973	186	1981	147	1989	144	1997	146
1974	167	1982	150	1990	145	1998	143
1975	162	1983	143	1991	144	1999	143
1976	167	1984	149	1992	144	2000	146
1977	168	1985	154	1993	145	2001	144
1978	165	1986	145	1994	142	2002	141
1979	155	1987	146	1995	145	2003	145

Make a graph that shows change over time. What overall pattern do you see? Have times stopped improving in recent years? If so, when did improvement end?

1.70 Household incomes Rich and poor households differ in ways that go beyond income. Figure 1.30 displays histograms that compare the distributions of household size (number of people) for low-income and high-income households in 2002.²⁵ Low-income households had incomes less than \$15,000, and high-income households had incomes of at least \$100,000.

Figure 1.30

The distributions of household size for households with incomes less than \$15,000 and households with incomes of at least \$100,000, for Exercise 1.70.



- About what percent of each group of households consisted of two people?
- What are the important differences between these two distributions? What do you think explains these differences?