DESCRIPTIVE STATISTICS

2.1 INTRODUCTION

In this chapter we introduce the subject matter of descriptive statistics, and in doing so learn ways to describe and summarize a set of data. Section 2.2 deals with ways of describing a data set. Subsections 2.2.1 and 2.2.2 indicate how data that take on only a relatively few distinct values can be described by using frequency tables or graphs, whereas Subsection 2.2.3 deals with data whose set of values is grouped into different intervals. Section 2.3 discusses ways of summarizing data sets by use of statistics, which are numerical quantities whose values are determined by the data. Subsection 2.3.1 considers three statistics that are used to indicate the "center" of the data set: the sample mean, the sample median, and the sample mode. Subsection 2.3.2 introduces the sample variance and its square root, called the sample standard deviation. These statistics are used to indicate the spread of the values in the data set. Subsection 2.3.3 deals with sample percentiles, which are statistics that tell us, for instance, which data value is greater than 95 percent of all the data. In Section 2.4 we present Chebyshev's inequality for sample data. This famous inequality gives an upper bound to the proportion of the data that can differ from the sample mean by more than k times the sample standard deviation. Whereas Chebyshev's inequality holds for all data sets, we can in certain situations, which are discussed in Section 2.5, obtain more precise estimates of the proportion of the data that is within k sample standard deviations of the sample mean. In Section 2.5 we note that when a graph of the data follows a bell-shaped form the data set is said to be approximately normal, and more precise estimates are given by the so-called empirical rule. Section 2.6 is concerned with situations in which the data consist of paired values. A graphical technique, called the scatter diagram, for presenting such data is introduced, as is the sample correlation coefficient, a statistic that indicates the degree to which a large value of the first member of the pair tends to go along with a large value of the second.

2.2 DESCRIBING DATA SETS

The numerical findings of a study should be presented clearly, concisely, and in such a manner that an observer can quickly obtain a feel for the essential characteristics of

the data. Over the years it has been found that tables and graphs are particularly useful ways of presenting data, often revealing important features such as the range, the degree of concentration, and the symmetry of the data. In this section we present some common graphical and tabular ways for presenting data.

2.2.1 Frequency Tables and Graphs

A data set having a relatively small number of distinct values can be conveniently presented in a *frequency table*. For instance, Table 2.1 is a frequency table for a data set consisting of the starting yearly salaries (to the nearest thousand dollars) of 42 recently graduated students with B.S. degrees in electrical engineering. Table 2.1 tells us, among other things, that the lowest starting salary of $57,000 was received by four of the graduates, whereas the highest salary of $70,000 was received by a single student. The most common starting salary was $62,000, and was received by 10 of the students.

TABLE 2.1 *Starting Yearly Salaries*

Starting Salary	Frequency
57	4
58	1
59	3
60	5
61	8
62	10
63	0
64	5
66	2
67	3
70	1

Data from a frequency table can be graphically represented by a *line graph* that plots the distinct data values on the horizontal axis and indicates their frequencies by the heights of vertical lines. A line graph of the data presented in Table 2.1 is shown in Figure 2.1.

When the lines in a line graph are given added thickness, the graph is called a *bar graph*. Figure 2.2 presents a bar graph.

Another type of graph used to represent a frequency table is the *frequency polygon*, which plots the frequencies of the different data values on the vertical axis, and then connects the plotted points with straight lines. Figure 2.3 presents a frequency polygon for the data of Table 2.1.

2.2.2 Relative Frequency Tables and Graphs

Consider a data set consisting of n values. If f is the frequency of a particular value, then the ratio f/n is called its *relative frequency*. That is, the relative frequency of a data value is

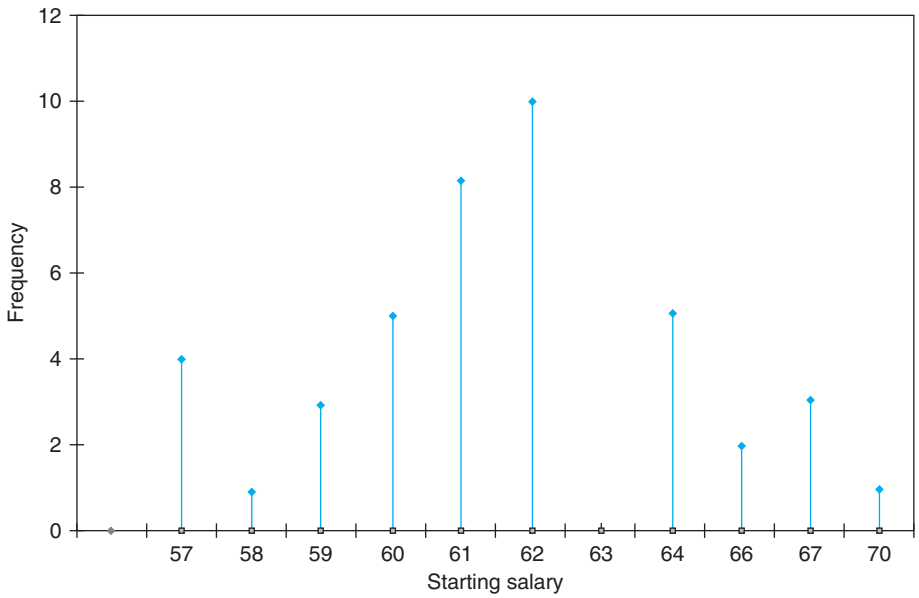

FIGURE 2.1 *Starting salary data.*

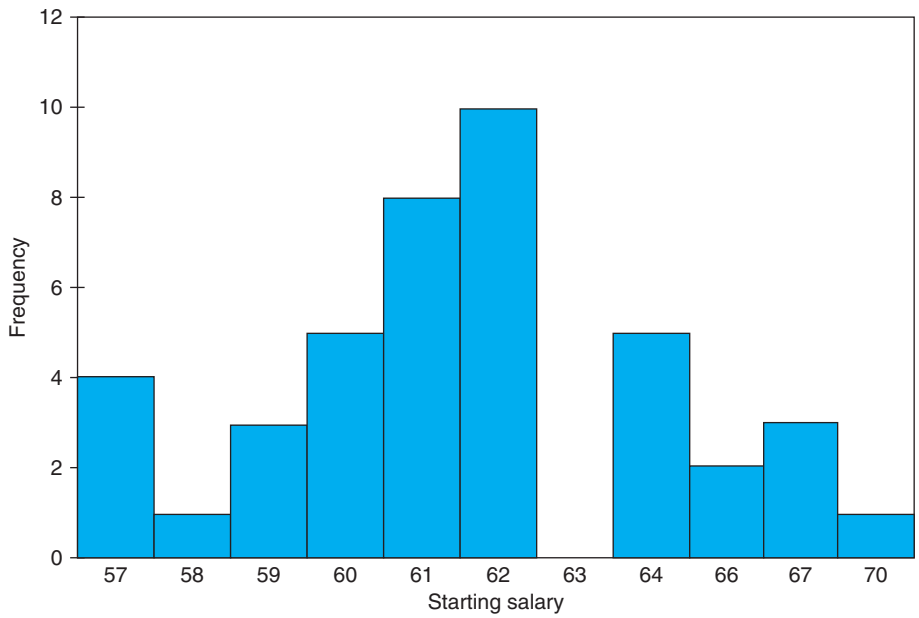

FIGURE 2.2 *Bar graph for starting salary data.*

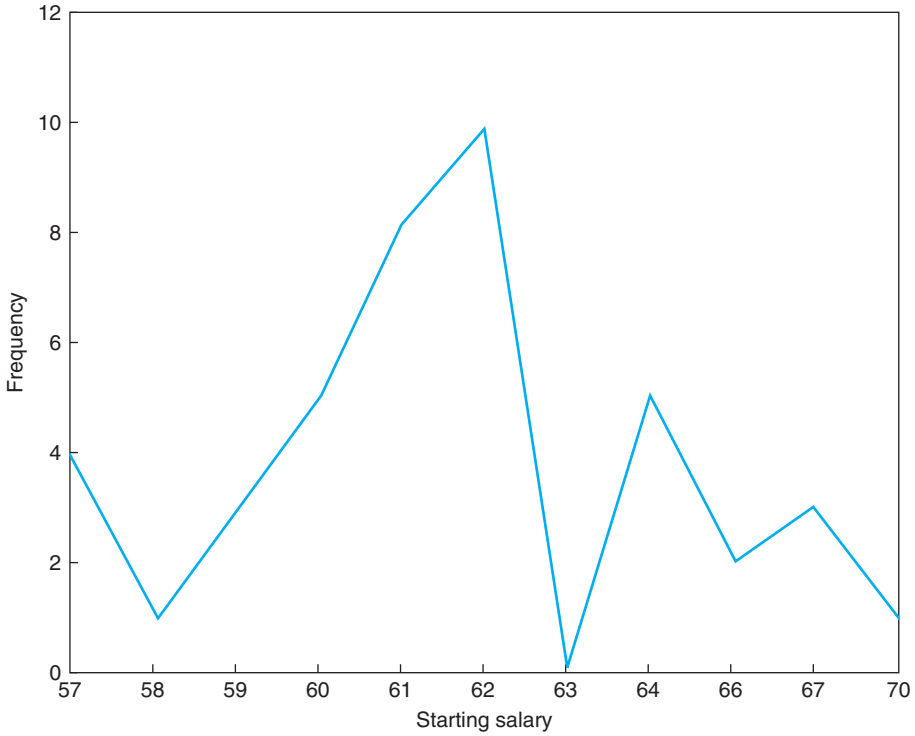

FIGURE 2.3 *Frequency polygon for starting salary data.*

the proportion of the data that have that value. The relative frequencies can be represented graphically by a relative frequency line or bar graph or by a relative frequency polygon. Indeed, these relative frequency graphs will look like the corresponding graphs of the absolute frequencies except that the labels on the vertical axis are now the old labels (that gave the frequencies) divided by the total number of data points.

EXAMPLE 2.2a Table 2.2 is a relative frequency table for the data of Table 2.1. The relative frequencies are obtained by dividing the corresponding frequencies of Table 2.1 by 42, the size of the data set. ■

A *pie chart* is often used to indicate relative frequencies when the data are not numerical in nature. A circle is constructed and then sliced into different sectors; one for each distinct type of data value. The relative frequency of a data value is indicated by the area of its sector, this area being equal to the total area of the circle multiplied by the relative frequency of the data value.

EXAMPLE 2.2b The following data relate to the different types of cancers affecting the 200 most recent patients to enroll at a clinic specializing in cancer. These data are represented in the pie chart presented in Figure 2.4. ■

TABLE 2.2

Starting Salary	Frequency
47	4/42 = .0952
48	1/42 = .0238
49	3/42
50	5/42
51	8/42
52	10/42
53	0
54	5/42
56	2/42
57	3/42
60	1/42

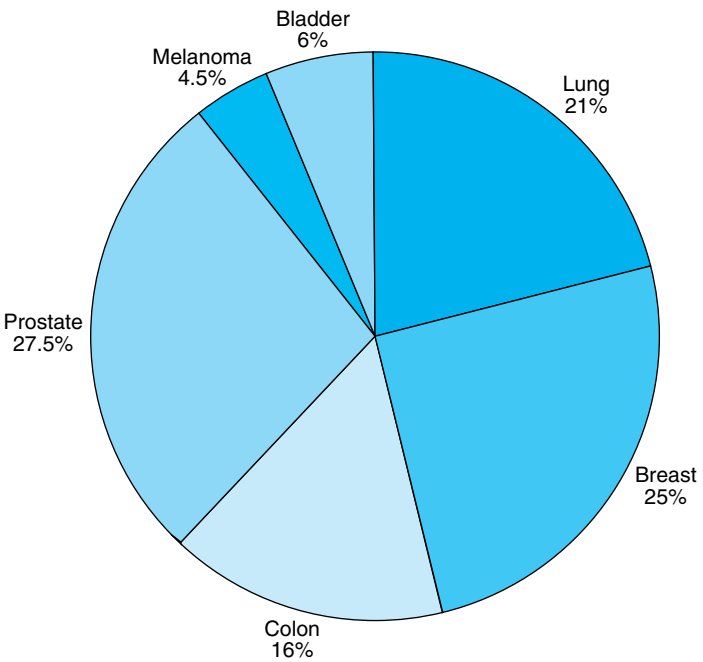

FIGURE 2.4

Type of Cancer	Number of New Cases	Relative Frequency
Lung	42	.21
Breast	50	.25
Colon	32	.16
Prostate	55	.275
Melanoma	9	.045
Bladder	12	.06

2.2.3 Grouped Data, Histograms, Ogives, and Stem and Leaf Plots

As seen in Subsection 2.2.2, using a line or a bar graph to plot the frequencies of data values is often an effective way of portraying a data set. However, for some data sets the number of distinct values is too large to utilize this approach. Instead, in such cases, it is useful to divide the values into groupings, or *class intervals*, and then plot the number of data values falling in each class interval. The number of class intervals chosen should be a trade-off between (1) choosing too few classes at a cost of losing too much information about the actual data values in a class and (2) choosing too many classes, which will result in the frequencies of each class being too small for a pattern to be discernible. Although 5 to 10

TABLE 2.3 *Life in Hours of 200 Incandescent Lamps*

Item Lifetimes									
1,067	919	1,196	785	1,126	936	918	1,156	920	948
855	1,092	1,162	1,170	929	950	905	972	1,035	1,045
1,157	1,195	1,195	1,340	1,122	938	970	1,237	956	1,102
1,022	978	832	1,009	1,157	1,151	1,009	765	958	902
923	1,333	811	1,217	1,085	896	958	1,311	1,037	702
521	933	928	1,153	946	858	1,071	1,069	830	1,063
930	807	954	1,063	1,002	909	1,077	1,021	1,062	1,157
999	932	1,035	944	1,049	940	1,122	1,115	833	1,320
901	1,324	818	1,250	1,203	1,078	890	1,303	1,011	1,102
996	780	900	1,106	704	621	854	1,178	1,138	951
1,187	1,067	1,118	1,037	958	760	1,101	949	992	966
824	653	980	935	878	934	910	1,058	730	980
844	814	1,103	1,000	788	1,143	935	1,069	1,170	1,067
1,037	1,151	863	990	1,035	1,112	931	970	932	904
1,026	1,147	883	867	990	1,258	1,192	922	1,150	1,091
1,039	1,083	1,040	1,289	699	1,083	880	1,029	658	912
1,023	984	856	924	801	1,122	1,292	1,116	880	1,173
1,134	932	938	1,078	1,180	1,106	1,184	954	824	529
998	996	1,133	765	775	1,105	1,081	1,171	705	1,425
610	916	1,001	895	709	860	1,110	1,149	972	1,002

class intervals are typical, the appropriate number is a subjective choice, and of course, you can try different numbers of class intervals to see which of the resulting charts appears to be most revealing about the data. It is common, although not essential, to choose class intervals of equal length.

The endpoints of a class interval are called the *class boundaries*. We will adopt the *left-end inclusion convention*, which stipulates that a class interval contains its left-end but not its right-end boundary point. Thus, for instance, the class interval 20–30 contains all values that are both greater than *or equal to* 20 and less than 30.

Table 2.3 presents the lifetimes of 200 incandescent lamps. A class frequency table for the data of Table 2.3 is presented in Table 2.4. The class intervals are of length 100, with the first one starting at 500.

TABLE 2.4 *A Class Frequency Table*

Class Interval	Frequency (Number of Data Values in the Interval)
500–600	2
600–700	5
700–800	12
800–900	25
900–1000	58
1000–1100	41
1100–1200	43
1200–1300	7
1300–1400	6
1400–1500	1

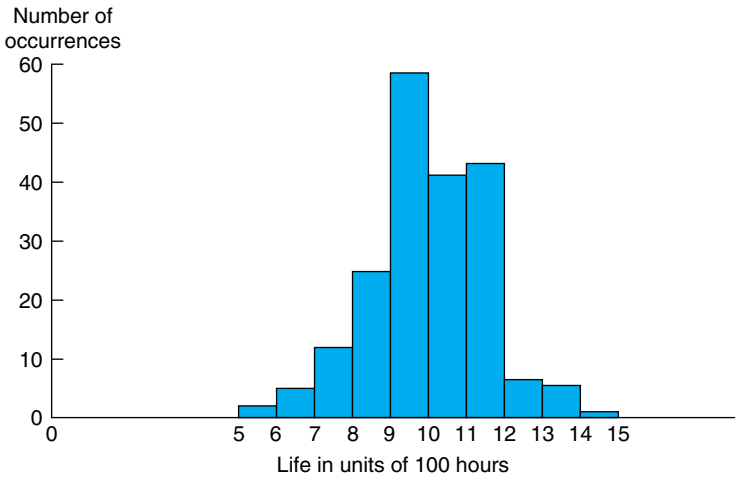

FIGURE 2.5 *A frequency histogram.*

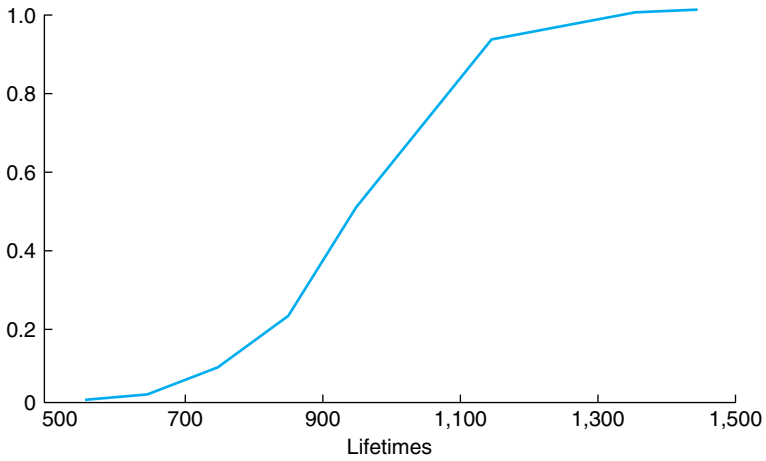

FIGURE 2.6 *A cumulative frequency plot.*

A bar graph plot of class data, with the bars placed adjacent to each other, is called a *histogram*. The vertical axis of a histogram can represent either the class frequency or the relative class frequency; in the former case the graph is called a *frequency histogram* and in the latter a *relative frequency histogram*. Figure 2.5 presents a frequency histogram of the data in Table 2.4.

We are sometimes interested in plotting a cumulative frequency (or cumulative relative frequency) graph. A point on the horizontal axis of such a graph represents a possible data value; its corresponding vertical plot gives the number (or proportion) of the data whose values are less than or equal to it. A cumulative relative frequency plot of the data of Table 2.3 is given in Figure 2.6. We can conclude from this figure that 100 percent of the data values are less than 1,500, approximately 40 percent are less than or equal to 900, approximately 80 percent are less than or equal to 1,100, and so on. A cumulative frequency plot is called an *ogive*.

An efficient way of organizing a small- to moderate-sized data set is to utilize a *stem and leaf plot*. Such a plot is obtained by first dividing each data value into two parts — its stem and its leaf. For instance, if the data are all two-digit numbers, then we could let the stem part of a data value be its tens digit and let the leaf be its ones digit. Thus, for instance, the value 62 is expressed as

Stem	**Leaf**
6	2

and the two data values 62 and 67 can be represented as

Stem	**Leaf**
6	2, 7

EXAMPLE 2.2c Table 2.5 gives the monthly and yearly average daily minimum temperatures in 35 U.S. cities.

The annual average daily minimum temperatures from Table 2.5 are represented in the following stem and leaf plot.

7	0.0
6	9.0
5	1.0, 1.3, 2.0, 5.5, 7.1, 7.4, 7.6, 8.5, 9.3
4	0.0, 1.0, 2.4, 3.6, 3.7, 4.8, 5.0, 5.2, 6.0, 6.7, 8.1, 9.0, 9.2
3	3.1, 4.1, 5.3, 5.8, 6.2, 9.0, 9.5, 9.5
2	9.0, 9.8 ■

2.3 SUMMARIZING DATA SETS

Modern-day experiments often deal with huge sets of data. For instance, in an attempt to learn about the health consequences of certain common practices, in 1951 the medical statisticians R. Doll and A. B. Hill sent questionnaires to all doctors in the United Kingdom and received approximately 40,000 replies. Their questions dealt with age, eating habits, and smoking habits. The respondents were then tracked for the ensuing 10 years and the causes of death for those who died were monitored. To obtain a feel for such a large amount of data, it is useful to be able to summarize it by some suitably chosen measures. In this section we present some summarizing *statistics*, where a statistic is a numerical quantity whose value is determined by the data.

2.3.1 Sample Mean, Sample Median, and Sample Mode

In this section we introduce some statistics that are used for describing the center of a set of data values. To begin, suppose that we have a data set consisting of the n numerical values $x_1, x_2, \ldots, x_n$. The sample mean is the arithmetic average of these values.

Definition

The *sample mean*, designated by $\bar{x}$, is defined by

$$\bar{x} = \sum_{i=1}^{n} x_i/n$$

The computation of the sample mean can often be simplified by noting that if for constants a and b

$$y_i = ax_i + b, \qquad i = 1, \ldots, n$$

TABLE 2.5 *Normal Daily Minimum Temperature — Selected Cities*

[In Fahrenheit degrees. Airport data except as noted. Based on standard 30-year period, 1961 through 1990]

State	Station	Jan.	Feb.	Mar.	Apr.	May	June	July	Aug.	Sept.	Oct.	Nov.	Dec.	Annual avg.
AL	Mobile	40.0	42.7	50.1	57.1	64.4	70.7	73.2	72.9	68.7	57.3	49.1	43.1	57.4
AK	Juneau	19.0	22.7	26.7	32.1	38.9	45.0	48.1	47.3	42.9	37.2	27.2	22.6	34.1
AZ	Phoenix	41.2	44.7	48.8	55.3	63.9	72.9	81.0	79.2	72.8	60.8	48.9	41.8	59.3
AR	Little Rock	29.1	33.2	42.2	50.7	59.0	67.4	71.5	69.8	63.5	50.9	41.5	33.1	51.0
CA	Los Angeles	47.8	49.3	50.5	52.8	56.3	59.5	62.8	64.2	63.2	59.2	52.8	47.9	55.5
	Sacramento	37.7	41.4	43.2	45.5	50.3	55.3	58.1	58.0	55.7	50.4	43.4	37.8	48.1
	San Diego	48.9	50.7	52.8	55.6	59.1	61.9	65.7	67.3	65.6	60.9	53.9	48.8	57.6
	San Francisco	41.8	45.0	45.8	47.2	49.7	52.6	53.9	55.0	55.2	51.8	47.1	42.7	49.0
CO	Denver	16.1	20.2	25.8	34.5	43.6	52.4	58.6	56.9	47.6	36.4	25.4	17.4	36.2
CT	Hartford	15.8	18.6	28.1	37.5	47.6	56.9	62.2	60.4	51.8	40.7	32.8	21.3	39.5
DE	Wilmington	22.4	24.8	33.1	41.8	52.2	61.6	67.1	65.9	58.2	45.7	37.0	27.6	44.8
DC	Washington	26.8	29.1	37.7	46.4	56.6	66.5	71.4	70.0	62.5	50.3	41.1	31.7	49.2
FL	Jacksonville	40.5	43.3	49.2	54.9	62.1	69.1	71.9	71.8	69.0	59.3	50.2	43.4	57.1
	Miami	59.2	60.4	64.2	67.8	72.1	75.1	76.2	76.7	75.9	72.1	66.7	61.5	69.0
GA	Atlanta	31.5	34.5	42.5	50.2	58.7	66.2	69.5	69.0	63.5	51.9	42.8	35.0	51.3
HI	Honolulu	65.6	65.4	67.2	68.7	70.3	72.2	73.5	74.2	73.5	72.3	70.3	67.0	70.0
ID	Boise	21.6	27.5	31.9	36.7	43.9	52.1	57.7	56.8	48.2	39.0	31.1	22.5	39.1
IL	Chicago	12.9	17.2	28.5	38.6	47.7	57.5	62.6	61.6	53.9	42.2	31.6	19.1	39.5
	Peoria	13.2	17.7	29.8	40.8	50.9	60.7	65.4	63.1	55.2	43.1	32.5	19.3	41.0
IN	Indianapolis	17.2	20.9	31.9	41.5	51.7	61.0	65.2	62.8	55.6	43.5	34.1	23.2	42.4
IA	Des Moines	10.7	15.6	27.6	40.0	51.5	61.2	66.5	63.6	54.5	42.7	29.9	16.1	40.0
KS	Wichita	19.2	23.7	33.6	44.5	54.3	64.6	69.9	67.9	59.2	46.6	33.9	23.0	45.0
KY	Louisville	23.2	26.5	36.2	45.4	54.7	62.9	67.3	65.8	58.7	45.8	37.3	28.6	46.0
LA	New Orleans	41.8	44.4	51.6	58.4	65.2	70.8	73.1	72.8	69.5	58.7	51.0	44.8	58.5
ME	Portland	11.4	13.5	24.5	34.1	43.4	52.1	58.3	57.1	48.9	38.3	30.4	17.8	35.8
MD	Baltimore	23.4	25.9	34.1	42.5	52.6	61.8	66.8	65.7	58.4	45.9	37.1	28.2	45.2
MA	Boston	21.6	23.0	31.3	40.2	49.8	59.1	65.1	64.0	56.8	46.9	38.3	26.7	43.6
MI	Detroit	15.6	17.6	27.0	36.8	47.1	56.3	61.3	59.6	52.5	40.9	32.2	21.4	39.0
	Sault Ste. Marie	4.6	4.8	15.3	28.4	38.4	45.5	51.3	51.3	44.3	36.2	25.9	11.8	29.8
MN	Duluth	−2.2	2.8	15.7	28.9	39.6	48.5	55.1	53.3	44.5	35.1	21.5	4.9	29.0
	Minneapolis-St. Paul	2.8	9.2	22.7	36.2	47.6	57.6	63.1	60.3	50.3	38.8	25.2	10.2	35.3
MS	Jackson	32.7	35.7	44.1	51.9	60.0	67.1	70.5	69.7	63.7	50.3	42.3	36.1	52.0
MO	Kansas City	16.7	21.8	32.6	43.8	53.9	63.1	68.2	65.7	56.9	45.7	33.6	21.9	43.7
	St. Louis	20.8	25.1	35.5	46.4	56.0	65.7	70.4	67.9	60.5	48.3	37.7	26.0	46.7
MT	Great Falls	11.6	17.2	22.8	31.9	40.9	48.6	53.2	52.2	43.5	35.8	24.3	14.6	33.1

Source: U.S. National Oceanic and Atmospheric Administration, Climatography of the United States, No. 81.

then the sample mean of the data set $y_1, \ldots, y_n$ is

$$\bar{y} = \sum_{i=1}^{n} (ax_i + b)/n = \sum_{i=1}^{n} ax_i/n + \sum_{i=1}^{n} b/n = a\bar{x} + b$$

EXAMPLE 2.3a The winning scores in the U.S. Masters golf tournament in the years from 2004 to 2013 were as follows:

$$280, 278, 272, 276, 281, 279, 276, 281, 289, 280$$

Find the sample mean of these scores.

SOLUTION Rather than directly adding these values, it is easier to first subtract 280 from each one to obtain the new values $y_i = x_i - 280$:

$$0, -2, -8, -4, 1, -1, -4, 1, 9, 0$$

Because the arithmetic average of the transformed data set is

$$\bar{y} = -8/10$$

it follows that

$$\bar{x} = \bar{y} + 280 = 279.2 \quad \blacksquare$$

Sometimes we want to determine the sample mean of a data set that is presented in a frequency table listing the k distinct values $v_1, \ldots, v_k$ having corresponding frequencies $f_1, \ldots, f_k$. Since such a data set consists of $n = \sum_{i=1}^{k} f_i$ observations, with the value v_i appearing f_i times, for each $i = 1, \ldots, k$, it follows that the sample mean of these n data values is

$$\bar{x} = \sum_{i=1}^{k} v_i f_i/n$$

By writing the preceding as

$$\bar{x} = \frac{f_1}{n} v_1 + \frac{f_2}{n} v_2 + \cdots + \frac{f_k}{n} v_k$$

we see that the sample mean is a *weighted average* of the distinct values, where the weight given to the value v_i is equal to the proportion of the n data values that are equal to v_i, $i = 1, \ldots, k$.

EXAMPLE 2.3b The following is a frequency table giving the ages of members of a symphony orchestra for young adults.

Age	Frequency
15	2
16	5
17	11
18	9
19	14
20	13

Find the sample mean of the ages of the 54 members of the symphony.

SOLUTION

$$\bar{x} = (15 \cdot 2 + 16 \cdot 5 + 17 \cdot 11 + 18 \cdot 9 + 19 \cdot 14 + 20 \cdot 13)/54 \approx 18.24 \quad \blacksquare$$

Another statistic used to indicate the center of a data set is the *sample median*; loosely speaking, it is the middle value when the data set is arranged in increasing order.

Definition

Order the values of a data set of size n from smallest to largest. If n is odd, the *sample median* is the value in position $(n+1)/2$; if n is even, it is the average of the values in positions $n/2$ and $n/2+1$.

Thus the sample median of a set of three values is the second smallest; of a set of four values, it is the average of the second and third smallest.

EXAMPLE 2.3c Find the sample median for the data described in Example 2.3b.

SOLUTION Since there are 54 data values, it follows that when the data are put in increasing order, the sample median is the average of the values in positions 27 and 28. Thus, the sample median is 18.5. $\blacksquare$

The sample mean and sample median are both useful statistics for describing the central tendency of a data set. The sample mean makes use of all the data values and is affected by extreme values that are much larger or smaller than the others; the sample median makes use of only one or two of the middle values and is thus not affected by extreme values. Which of them is more useful depends on what one is trying to learn from the data. For instance, if a city government has a flat rate income tax and is trying to estimate its total revenue from the tax, then the sample mean of its residents' income would be a more useful statistic. On the other hand, if the city was thinking about constructing middle-income housing, and wanted to determine the proportion of its population able to afford it, then the sample median would probably be more useful.

EXAMPLE 2.3d In a study reported in Hoel, D. G., "A representation of mortality data by competing risks," *Biometrics*, **28**, pp. 475–488, 1972, a group of 5-week-old mice were each given a radiation dose of 300 rad. The mice were then divided into two groups; the first group was kept in a germ-free environment, and the second in conventional laboratory conditions. The numbers of days until death were then observed. The data for those whose death was due to thymic lymphoma are given in the following stem and leaf plots (whose stems are in units of hundreds of days); the first plot is for mice living in the germ-free conditions and the second for mice living under ordinary laboratory conditions.

Germ-Free Mice

1	58, 92, 93, 94, 95
2	02, 12, 15, 29, 30, 37, 40, 44, 47, 59
3	01, 01, 21, 37
4	15, 34, 44, 85, 96
5	29, 37
6	24
7	07
8	00

Conventional Mice

1	59, 89, 91, 98
2	35, 45, 50, 56, 61, 65, 66, 80
3	43, 56, 83
4	03, 14, 28, 32

Determine the sample means and the sample medians for the two sets of mice.

SOLUTION It is clear from the stem and leaf plots that the sample mean for the set of mice put in the germ-free setting is larger than the sample mean for the set of mice in the usual laboratory setting; indeed, a calculation gives that the former sample mean is 344.07, whereas the latter one is 292.32. On the other hand, since there are 29 data values for the germ-free mice, the sample median is the 15th largest data value, namely, 259; similarly, the sample median for the other set of mice is the 10th largest data value, namely, 265. Thus, whereas the sample mean is quite a bit larger for the first data set, the sample medians are approximately equal. The reason for this is that whereas the sample mean for the first set is greatly affected by the five data values greater than 500, these values have a much smaller effect on the sample median. Indeed, the sample median would remain unchanged if these values were replaced by any other five values greater than or equal to 259. It appears from the stem and leaf plots that the germ-free conditions probably improved the life span of the five longest living rats, but it is unclear what, if any, effect it had on the life spans of the other rats. ■

Another statistic that has been used to indicate the central tendency of a data set is the *sample mode*, defined to be the value that occurs with the greatest frequency. If no single value occurs most frequently, then all the values that occur at the highest frequency are called *modal values*.

EXAMPLE 2.3e The following frequency table gives the values obtained in 40 rolls of a die.

Value	Frequency
1	9
2	8
3	5
4	5
5	6
6	7

Find **(a)** the sample mean, **(b)** the sample median, and **(c)** the sample mode.

SOLUTION **(a)** The sample mean is

$$\bar{x} = (9 + 16 + 15 + 20 + 30 + 42)/40 = 3.05$$

(b) The sample median is the average of the 20th and 21st smallest values, and is thus equal to 3. **(c)** The sample mode is 1, the value that occurred most frequently. ■

2.3.2 Sample Variance and Sample Standard Deviation

Whereas we have presented statistics that describe the central tendencies of a data set, we are also interested in ones that describe the spread or variability of the data values. A statistic that could be used for this purpose would be one that measures the average value of the squares of the distances between the data values and the sample mean. This is accomplished by the sample variance, which for technical reasons divides the sum of the squares of the differences by $n - 1$ rather than n, where n is the size of the data set.

Definition

The *sample variance*, call it s^2, of the data set $x_1, \ldots, x_n$ is defined by

$$s^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2/(n-1)$$

EXAMPLE 2.3f Find the sample variances of the data sets **A** and **B** given below.

$$\mathbf{A}: 3, 4, 6, 7, 10 \qquad \mathbf{B}: -20, 5, 15, 24$$

SOLUTION As the sample mean for data set **A** is $\bar{x} = (3+4+6+7+10)/5 = 6$, it follows that its sample variance is

$$s^2 = [(-3)^2 + (-2)^2 + 0^2 + 1^2 + 4^2]/4 = 7.5$$

The sample mean for data set **B** is also 6; its sample variance is

$$s^2 = [(-26)^2 + (-1)^2 + 9^2 + (18)^2]/3 \approx 360.67$$

Thus, although both data sets have the same sample mean, there is a much greater variability in the values of the **B** set than in the **A** set. ■

The following algebraic identity is often useful for computing the sample variance:

An Algebraic Identity

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$$

The identity is proven as follows:

$$\begin{aligned}\sum_{i=1}^{n}(x_i - \bar{x})^2 &= \sum_{i=1}^{n}\left(x_i^2 - 2x_i\bar{x} + \bar{x}^2\right)\\ &= \sum_{i=1}^{n} x_i^2 - 2\bar{x}\sum_{i=1}^{n} x_i + \sum_{i=1}^{n}\bar{x}^2\\ &= \sum_{i=1}^{n} x_i^2 - 2n\bar{x}^2 + n\bar{x}^2\\ &= \sum_{i=1}^{n} x_i^2 - n\bar{x}^2\end{aligned}$$

The computation of the sample variance can also be eased by noting that if

$$y_i = a + bx_i, \qquad i = 1, \ldots, n$$

then $\bar{y} = a + b\bar{x}$, and so

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = b^2\sum_{i=1}^{n}(x_i - \bar{x})^2$$

That is, if s_y^2 and s_x^2 are the respective sample variances, then

$$s_y^2 = b^2 s_x^2$$

In other words, adding a constant to each data value does not change the sample variance; whereas multiplying each data value by a constant results in a new sample variance that is equal to the old one multiplied by the square of the constant. ■

EXAMPLE 2.3g The following data give the worldwide number of fatal airline accidents of commercially scheduled air transports in the years from 1997 to 2005.

Year	1997	1998	1999	2000	2001	2002	2003	2004	2005
Accidents	25	20	21	18	13	13	7	9	18

Source: National Safety Council.

Find the sample variance of the number of accidents in these years.

SOLUTION Let us start by subtracting 18 from each value, to obtain the new data set:

$$7, 2, 3, 0, -5, -5, -11, -9, 0$$

Calling the transformed data $y_1, \ldots, y_9$, we have

$$\bar{y} = \sum_{i=1}^{9} y_i/9 = -2, \qquad \sum_{i=1}^{n} y_i^2 = 49 + 4 + 9 + 25 + 25 + 121 + 81 = 314$$

Hence, since the sample variance of the transformed data is equal to that of the original data, upon using the algebraic identity we obtain

$$s^2 = \frac{314 - 9(4)}{8} = 34.75 \quad ■$$

Program 2.3 on the text disk can be used to obtain the sample variance for large data sets.

The positive square root of the sample variance is called the *sample standard deviation*.

Definition

The quantity s, defined by

$$s = \sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2/(n-1)}$$

is called the *sample standard deviation*.

The sample standard deviation is measured in the same units as the data.

2.3.3 Sample Percentiles and Box Plots

Loosely speaking, the sample $100p$ percentile of a data set is that value such that $100p$ percent of the data values are less than or equal to it, $0 \leq p \leq 1$. More formally, we have the following definition.

Definition

The *sample 100*p *percentile* is that data value such that at least $100p$ percent of the data are less than or equal to it and at least $100(1-p)$ percent are greater than or equal to it. If two data values satisfy this condition, then the sample $100p$ percentile is the arithmetic average of these two values.

To determine the sample $100p$ percentile of a data set of size n, we need to determine the data values such that

1. At least np of the values are less than or equal to it.
2. At least $n(1-p)$ of the values are greater than or equal to it.

To accomplish this, first arrange the data in increasing order. Then, note that if np is not an integer, then the only data value that satisfies the preceding conditions is the one whose position when the data are ordered from smallest to largest is the smallest integer exceeding np. For instance, if $n=22$, $p=.8$, then we require a data value such that at least 17.6 of the values are less than or equal to it, and at least 4.4 of them are greater than or equal to it. Clearly, only the 18th smallest value satisfies both conditions and this is the sample 80 percentile. On the other hand, if np is an integer, then it is easy to check that both the values in positions np and $np+1$ satisfy the preceding conditions, and so the sample $100p$ percentile is the average of these values. For instance, if we wanted the 90 percentile of a data set of size 20, then both the (18)th and (19)th smallest values would be such that at least 90 percent of the data values are less than or equal to them, and at least 10 percent of the data values are greater than or equal to them. Thus, the 90 percentile is the average of these two values.

EXAMPLE 2.3h Table 2.6 lists the populations of the 25 most populous U.S. cities for the year 1994. For this data set, find **(a)** the sample 10 percentile and **(b)** the sample 80 percentile.

SOLUTION **(a)** Because the sample size is 25 and $25(.10)=2.5$, the sample 10 percentile is the third smallest value, equal to 590,763.

(b) Because $25(.80)=20$, the sample 80 percentile is the average of the twentieth and the twenty-first smallest values. Hence, the sample 80 percentile is

$$\frac{1{,}512{,}986+1{,}448{,}394}{2}=1{,}480{,}690 \quad \blacksquare$$

The sample 50 percentile is, of course, just the sample median. Along with the sample 25 and 75 percentiles, it makes up the sample quartiles.

Definition

The sample 25 percentile is called the *first quartile*; the sample 50 percentile is called the sample median or the *second quartile*; the sample 75 percentile is called the *third quartile*.

TABLE 2.6 *Population of 25 Largest U.S. Cities, July 2006*

Rank	City	Population
1	New York, NY	8,250,567
2	Los Angeles, CA	3,849,378
3	Chicago, IL	2,833,321
4	Houston, TX	2,144,491
5	Phoenix, AR	1,512,986
6	Philadelphia, PA	1,448,394
7	San Antonio, TX	1,296,682
8	San Diego, CA	1,256,951
9	Dallas, TX	1,232,940
10	San Jose, CA	929,936
11	Detroit, MI	918,849
12	Jacksonville, FL	794,555
13	Indianapolis, IN	785,597
14	San Francisco, CA	744,041
15	Columbus, OH	733,203
16	Austin, TX	709,893
17	Memphis, TN	670,902
18	Fort Worth, TX	653,320
19	Baltimore, MD	640,961
20	Charlotte, NC	630,478
21	El Paso, TX	609,415
22	Milwaukee, WI	602,782
23	Boston, MA	590,763
24	Seattle, WA	582,454
25	Washington, DC	581,530

The quartiles break up a data set into four parts, with roughly 25 percent of the data being less than the first quartile, 25 percent being between the first and second quartile, 25 percent being between the second and third quartile, and 25 percent being greater than the third quartile.

EXAMPLE 2.3i Noise is measured in decibels, denoted as dB. One decibel is about the level of the weakest sound that can be heard in a quiet surrounding by someone with good hearing; a whisper measures about 30 dB; a human voice in normal conversation is about 70 dB; a loud radio is about 100 dB. Ear discomfort usually occurs at a noise level of about 120 dB.

The following data give noise levels measured at 36 different times directly outside of Grand Central Station in Manhattan.

82, 89, 94, 110, 74, 122, 112, 95, 100, 78, 65, 60, 90, 83, 87, 75, 114, 85
69, 94, 124, 115, 107, 88, 97, 74, 72, 68, 83, 91, 90, 102, 77, 125, 108, 65

Determine the quartiles.

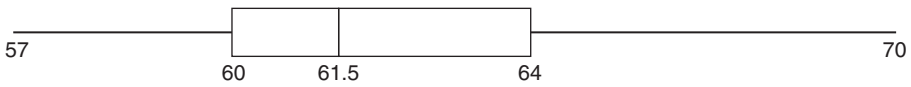

FIGURE 2.7 *A box plot.*

SOLUTION A stem and leaf plot of the data is as follows:

6	0, 5, 5, 8, 9
7	2, 4, 4, 5, 7, 8
8	2, 3, 3, 5, 7, 8, 9
9	0, 0, 1, 4, 4, 5, 7
10	0, 2, 7, 8
11	0, 2, 4, 5
12	2, 4, 5

Because $36/4 = 9$, the first quartile is 74.5, the average of the 9th and 10th smallest data values; the second quartile is 89.5, the average of the 18th and 19th smallest values; the third quartile is 104.5, the average of the 27th and 28th smallest values. ■

A *box plot* is often used to plot some of the summarizing statistics of a data set. A straight line segment stretching from the smallest to the largest data value is drawn on a horizontal axis; imposed on the line is a "box," which starts at the first and continues to the third quartile, with the value of the second quartile indicated by a vertical line. For instance, the 42 data values presented in Table 2.1 go from a low value of 57 to a high value of 70. The value of the first quartile (equal to the value of the 11th smallest on the list) is 60; the value of the second quartile (equal to the average of the 21st and 22nd smallest values) is 61.5; and the value of the third quartile (equal to the value of the 32nd smallest on the list) is 64. The box plot for this data set is shown in Figure 2.7.

The length of the line segment on the box plot, equal to the largest minus the smallest data value, is called the *range* of the data. Also, the length of the box itself, equal to the third quartile minus the first quartile, is called the *interquartile range*.

2.4 CHEBYSHEV'S INEQUALITY

Let $\bar{x}$ and s be the sample mean and sample standard deviation of a data set. Assuming that $s > 0$, Chebyshev's inequality states that for any value of $k \geq 1$, greater than $100(1 - 1/k^2)$ percent of the data lie within the interval from $\bar{x} - ks$ to $\bar{x} + ks$. Thus, by letting $k = 3/2$, we obtain from Chebyshev's inequality that greater than $100(5/9) = 55.56$ percent of the data from any data set lies within a distance $1.5s$ of the sample mean $\bar{x}$; letting $k = 2$ shows that greater than 75 percent of the data lies within $2s$ of the sample mean; and letting $k = 3$ shows that greater than $800/9 \approx 88.9$ percent of the data lies within 3 sample standard deviations of $\bar{x}$.

When the size of the data set is specified, Chebyshev's inequality can be sharpened, as indicated in the following formal statement and proof.

Chebyshev's Inequality

Let $\bar{x}$ and s be the sample mean and sample standard deviation of the data set consisting of the data $x_1, \ldots, x_n$, where $s > 0$. Let

$$S_k = \{i, 1 \le i \le n : |x_i - \bar{x}| < ks\}$$

and let $|S_k|$ be the number of elements in the set S_k. Then, for any $k \ge 1$,

$$\frac{|S_k|}{n} \ge 1 - \frac{n-1}{nk^2} > 1 - \frac{1}{k_2}$$

Proof

$$\begin{aligned}
(n-1)s^2 &= \sum_{i=1}^{n} (x_i - \bar{x})^2 \\
&= \sum_{i \in S_k} (x_i - \bar{x})^2 + \sum_{i \notin S_k} (x_i - \bar{x})^2 \\
&\ge \sum_{i \notin S_k} (x_i - \bar{x})^2 \\
&\ge \sum_{i \notin S_k} k^2 s^2 \\
&= k^2 s^2 (n - |S_k|)
\end{aligned}$$

where the first inequality follows because all terms being summed are nonnegative, and the second follows since $(x_1 - \bar{x})^2 \ge k^2 s^2$ when $i \notin S_k$. Dividing both sides of the preceding inequality by $nk^2 s^2$ yields that

$$\frac{n-1}{nk^2} \ge \frac{n - |S_k|}{n} = 1 - \frac{|S_k|}{n}$$

and the result is proven. ■

Because Chebyshev's inequality holds universally, it might be expected for given data that the actual percentage of the data values that lie within the interval from $\bar{x} - ks$ to $\bar{x} + ks$ might be quite a bit larger than the bound given by the inequality.

EXAMPLE 2.4a Table 2.7 lists the 10 top-selling passenger cars in the United States in the month of June, 2013.

TABLE 2.7 *Top Selling Vehicles*

June 2013 Sales (in thousands of vehicles)	
Ford F Series	68.0
Chevrolet Silverado	43.3
Toyoto Camry	35.9
Chevrolet Cruze	32.9
Honda Accord	31.7
Honda Civic	29.7
Dodge Ram	29.6
Ford Escape	28.7
Nissan Altima	26.9
Honda CR – V	26.6

A simple calculation yields that the sample mean and sample standard deviation of these data are

$$\bar{x} = 35.33 \qquad s = 11.86$$

Thus Chebyshev's Inequality states that at least $100(5/9) = 55.55$ percent of the data lies in the interval

$$\left(\bar{x} - \frac{3}{2}s, \quad \bar{x} + \frac{3}{2}s\right) = (17.54, \quad 53.12)$$

whereas, in actuality, 90 percent of the data falls within these limits. ■

Suppose now that we are interested in the fraction of data values that exceed the sample mean by at least k sample standard deviations, where k is positive. That is, suppose that $\bar{x}$ and s are the sample mean and the sample standard deviation of the data set $x_1, x_2, \ldots, x_n$. Then, with

$$N(k) = \text{number of } i : x_i - \bar{x} \geq ks$$

what can we say about $N(k)/n$? Clearly,

$$\begin{aligned} \frac{N(k)}{n} &\leq \frac{\text{number of } i : |x_i - \bar{x}| \geq ks}{n} \\ &\leq \frac{1}{k^2} \quad \text{by Chebyshev's inequality} \end{aligned}$$

However, we can make a stronger statement, as is shown in the following one-sided version of Chebyshev's inequality.

The One-Sided Chebyshev Inequality

Let $\bar{x}$ and s be the sample mean and sample standard deviation of the data set consisting of the data $x_1, \ldots, x_n$. Suppose $s > 0$, and let $N(k) = \text{number of } i : x_i - \bar{x} \geq ks$. Then, for any $k > 0$,

$$\frac{N(k)}{n} \leq \frac{1}{1+k^2}$$

Proof

Let $y_i = x_i - \bar{x},\ i = 1, \ldots, n$. For any $b > 0$, we have that

$$\begin{aligned}\sum_{i=1}^{n} (y_i + b)^2 &\geq \sum_{i:y_i \geq ks} (y_i + b)^2 \\ &\geq \sum_{i:y_i \geq ks} (ks + b)^2 \\ &= N(k)(ks + b)^2 \end{aligned} \tag{2.4.1}$$

where the first inequality follows because $(y_i + b)^2 \geq 0$, the second because both ks and b are positive, and the final equality because $N(k)$ is equal to the number of i such that $y_i \geq ks$. However,

$$\begin{aligned}\sum_{i=1}^{n} (y_i + b)^2 &= \sum_{i=1}^{n} (y_i^2 + 2by_i + b^2) \\ &= \sum_{i=1}^{n} y_i^2 + 2b\sum_{i=1}^{n} y_i + nb^2 \\ &= \sum_{i=1}^{n} y_i^2 + nb^2 \\ &= (n-1)s^2 + nb^2 \end{aligned}$$

where the next to last equation used that $\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} (x_i - \bar{x}) = \sum_{i=1}^{n} x_i - n\bar{x} = 0$. Therefore, we obtain from Equation (2.4.1) that

$$N(k) \leq \frac{(n-1)s^2 + nb^2}{(ks + b)^2} < \frac{ns^2 + nb^2}{(ks + b)^2}$$

implying that

$$\frac{N(k)}{n} \leq \frac{s^2 + b^2}{(ks + b)^2}$$

Because the preceding is valid for all $b > 0$, we can set $b = s/k$ (which is the value of b that minimizes the right-hand side of the preceding) to obtain that

$$\frac{N(k)}{n} \leq \frac{s^2 + s^2/k^2}{(ks + s/k)^2}$$

Multiplying the numerator and the denominator of the right side of the preceding by k^2/s^2 gives

$$\frac{N(k)}{n} \leq \frac{k^2+1}{(k^2+1)^2} = \frac{1}{k^2+1}$$

and the result is proven. Thus, for instance, where the usual Chebyshev inequality shows that at most 25 percent of data values are at least 2 standard deviations greater than the sample mean, the one-sided Chebyshev inequality lowers the bound to "at most 20 percent." ■

2.5 NORMAL DATA SETS

Many of the large data sets observed in practice have histograms that are similar in shape. These histograms often reach their peaks at the sample median and then decrease on both sides of this point in a bell-shaped symmetric fashion. Such data sets are said to be *normal* and their histograms are called *normal histograms*. Figure 2.8 is the histogram of a normal data set.

If the histogram of a data set is close to being a normal histogram, then we say that the data set is *approximately normal*. For instance, we would say that the histogram given in Figure 2.9 is from an approximately normal data set, whereas the ones presented in Figures 2.10 and 2.11 are not (because each is too nonsymmetric). Any data set that is not approximately symmetric about its sample median is said to be *skewed*. It is "skewed to the right" if it has a long tail to the right and "skewed to the left" if it has a long tail to the left. Thus the data set presented in Figure 2.10 is skewed to the left and the one of Figure 2.11 is skewed to the right.

It follows from the symmetry of the normal histogram that a data set that is approximately normal will have its sample mean and sample median approximately equal.

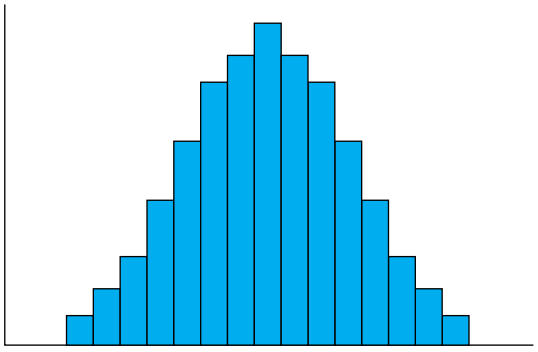

FIGURE 2.8 *Histogram of a normal data set.*

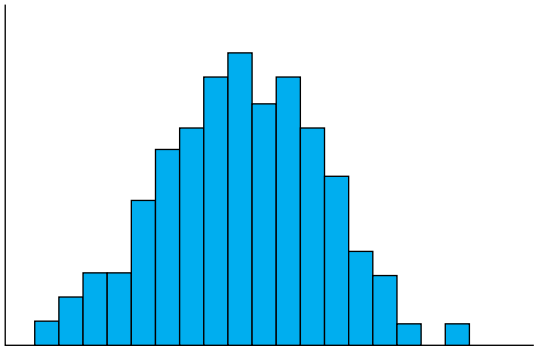

FIGURE 2.9 *Histogram of an approximately normal data set.*

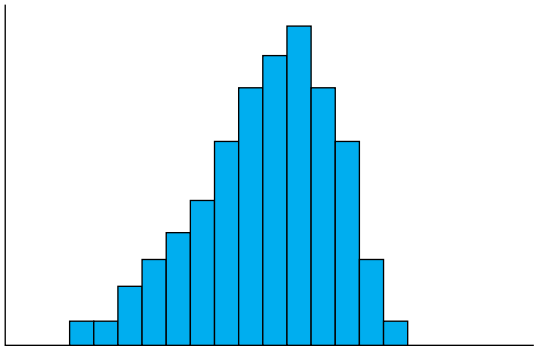

FIGURE 2.10 *Histogram of a data set skewed to the left.*

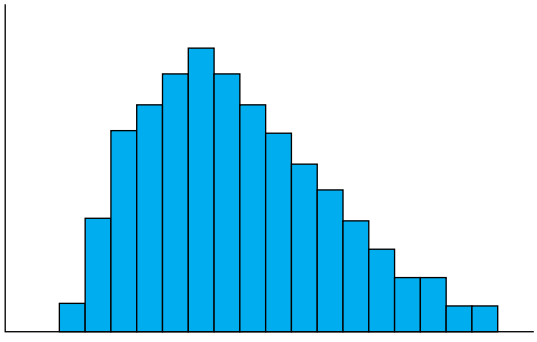

FIGURE 2.11 *Histogram of a data set skewed to the right.*

Suppose that $\bar{x}$ and s are the sample mean and sample standard deviation of an approximately normal data set. The following rule, known as the *empirical rule*, specifies the approximate proportions of the data observations that are within s, $2s$, and $3s$ of the sample mean $\bar{x}$.

The Empirical Rule

If a data set is approximately normal with sample mean $\bar{x}$ and sample standard deviation s, then the following statements are true.

1. Approximately 68 percent of the observations lie within
$$\bar{x} \pm s$$
2. Approximately 95 percent of the observations lie within
$$\bar{x} \pm 2s$$
3. Approximately 99.7 percent of the observations lie within
$$\bar{x} \pm 3s$$

EXAMPLE 2.5a The following stem and leaf plot gives the scores on a statistics exam taken by industrial engineering students.

9	0, 1, 4
8	3, 5, 5, 7, 8
7	2, 4, 4, 5, 7, 7, 8
6	0, 2, 3, 4, 6, 6
5	2, 5, 5, 6, 8
4	3, 6

By standing the stem and leaf plot on its side we can see that the corresponding histogram is approximately normal. Use it to assess the empirical rule.

SOLUTION A calculation gives that

$$\bar{x} \approx 70.571, \quad s \approx 14.354$$

Thus the empirical rule states that approximately 68 percent of the data are between 56.2 and 84.9; the actual percentage is $1{,}500/28 \approx 53.6$. Similarly, the empirical rule gives that approximately 95 percent of the data are between 41.86 and 99.28, whereas the actual percentage is 100. ■

A data set that is obtained by sampling from a population that is itself made up of subpopulations of different types is usually not normal. Rather, the histogram from such a data set often appears to resemble a combining, or superposition, of normal histograms

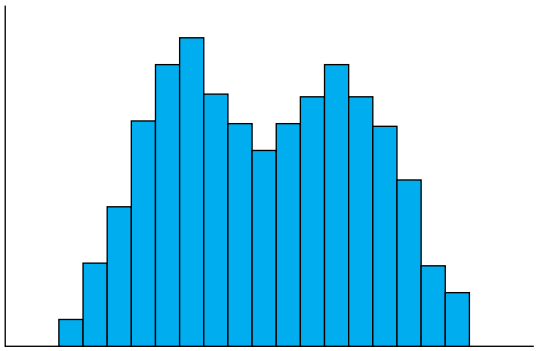

FIGURE 2.12 *Histogram of a bimodal data set.*

and thus will often have more than one local peak or hump. Because the histogram will be higher at these local peaks than at their neighboring values, these peaks are similar to modes. A data set whose histogram has two local peaks is said to be *bimodal*. The data set represented in Figure 2.12 is bimodal.

2.6 PAIRED DATA SETS AND THE SAMPLE CORRELATION COEFFICIENT

We are often concerned with data sets that consist of pairs of values that have some relationship to each other. If each element in such a data set has an x value and a y value, then we represent the ith data point by the pair (x_i, y_i). For instance, in an attempt to determine the relationship between the daily midday temperature (measured in degrees Celsius) and the number of defective parts produced during that day, a company recorded the data presented in Table 2.8. For this data set, x_i represents the temperature in degrees Celsius and y_i the number of defective parts produced on day i.

A useful way of portraying a data set of paired values is to plot the data on a two-dimensional graph, with the x-axis representing the x value of the data and the y-axis representing the y value. Such a plot is called a *scatter diagram*. Figure 2.13 presents a scatter diagram for the data of Table 2.8.

A question of interest concerning paired data sets is whether large x values tend to be paired with large y values, and small x values with small y values; if this is not the case, then we might question whether large values of one of the variables tend to be paired with small values of the other. A rough answer to these questions can often be provided by the scatter diagram. For instance, Figure 2.13 indicates that there appears to be some connection between high temperatures and large numbers of defective items. To obtain a quantitative measure of this relationship, we now develop a statistic that attempts to measure the degree to which larger x values go with larger y values and smaller x values with smaller y values.

TABLE 2.8 *Temperature and Defect Data*

Day	Temperature	Number of Defects
1	24.2	25
2	22.7	31
3	30.5	36
4	28.6	33
5	25.5	19
6	32.0	24
7	28.6	27
8	26.5	25
9	25.3	16
10	26.0	14
11	24.4	22
12	24.8	23
13	20.6	20
14	25.1	25
15	21.4	25
16	23.7	23
17	23.9	27
18	25.2	30
19	27.4	33
20	28.3	32
21	28.8	35
22	26.6	24

Suppose that the data set consists of the paired values (x_i, y_i), $i = 1, \ldots, n$. To obtain a statistic that can be used to measure the association between the individual values of a set of paired data, let $\bar{x}$ and $\bar{y}$ denote the sample means of the x values and the y values, respectively. For data pair i, consider $x_i - \bar{x}$ the deviation of its x value from the sample mean, and $y_i - \bar{y}$ the deviation of its y value from the sample mean. Now if x_i is a large x value, then it will be larger than the average value of all the x's, so the deviation $x_i - \bar{x}$ will be a positive value. Similarly, when x_i is a small x value, then the deviation $x_i - \bar{x}$ will be a negative value. Because the same statements are true about the y deviations, we can conclude the following:

> When large values of the x variable tend to be associated with large values of the y variable and small values of the x variable tend to be associated with small values of the y variable, then the signs, either positive or negative, of $x_i - \bar{x}$ and $y_i - \bar{y}$ will tend to be the same.

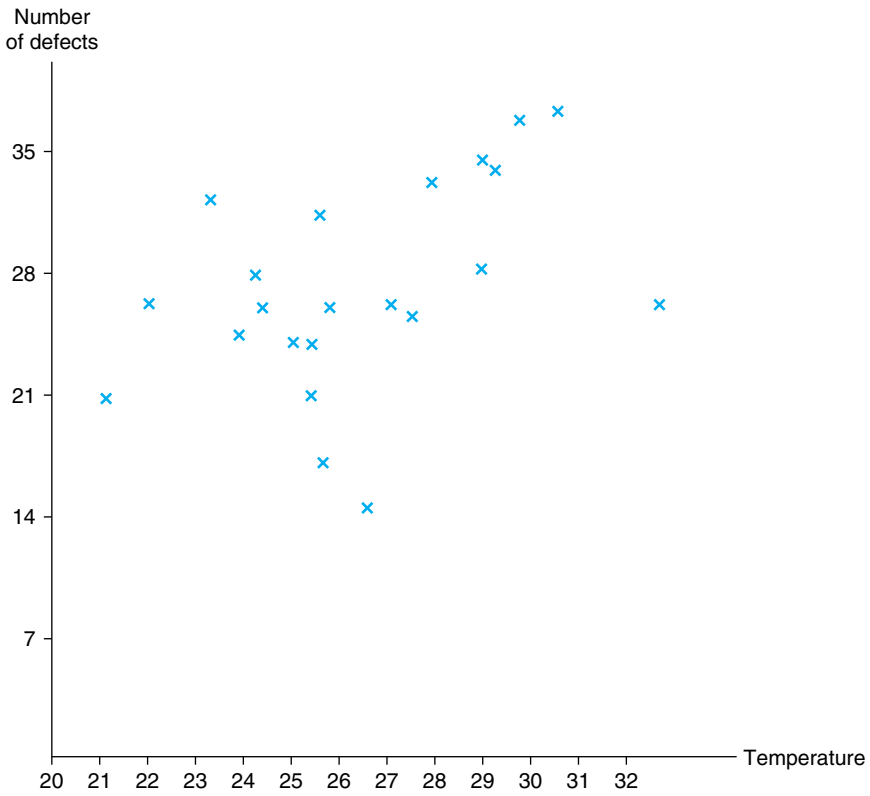

FIGURE 2.13 *A scatter diagram.*

Now, if $x_i - \bar{x}$ and $y_i - \bar{y}$ both have the same sign (either positive or negative), then their product $(x_i - \bar{x})(y_i - \bar{y})$ will be positive. Thus, it follows that when large x values tend to be associated with large y values and small x values are associated with small y values, then $\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$ will tend to be a large positive number. [In fact, not only will all the products have a positive sign when large (small) x values are paired with large (small) y values, but it also follows from a mathematical result known as Hardy's lemma that the largest possible value of the sum of paired products will be obtained when the largest $x_i - \bar{x}$ is paired with the largest $y_i - \bar{y}$, the second largest $x_i - \bar{x}$ is paired with the second largest $y_i - \bar{y}$, and so on.] In addition, it similarly follows that when large values of x_i tend to be paired with small values of y_i then the signs of $x_i - \bar{x}$ and $y_i - \bar{y}$ will be opposite and so $\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$ will be a large negative number.

To determine what it means for $\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$ to be "large," we standardize this sum first by dividing by $n - 1$ and then by dividing by the product of the two sample standard deviations. The resulting statistic is called the *sample correlation coefficient.*

Definition

Consider the data pairs (x_i, y_i), $i = 1, \ldots, n$. and let s_x and s_y denote, respectively, the sample standard deviations of the x values and the y values. The *sample correlation coefficient,* call it r, of the data pairs (x_i, y_i), $i = 1, \ldots, n$ is defined by

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}$$

$$= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

When $r > 0$ we say that the sample data pairs are *positively correlated*, and when $r < 0$ we say that they are *negatively correlated.*

The following are properties of the sample correlation coefficient.

Properties of *r*

1. $$-1 \le r \le 1$$
2. If for constants a and b, with $b > 0$,

   $$y_i = a + bx_i, \qquad i = 1, \ldots, n$$

   then $r = 1$.
3. If for constants a and b, with $b < 0$,

   $$y_i = a + bx_i, \qquad i = 1, \ldots, n$$

   then $r = -1$.
4. If r is the sample correlation coefficient for the data pairs x_i, y_i, $i = 1, \ldots, n$ then it is also the sample correlation coefficient for the data pairs

   $$a + bx_i, \quad c + dy_i, \quad i = 1, \ldots, n$$

   provided that b and d are both positive or both negative.

Property 1 says that the sample correlation coefficient r is always between -1 and $+1$. Property 2 says that r will equal $+1$ when there is a straight line (also called a linear) relation between the paired data such that large y values are attached to large x values. Property 3 says that r will equal -1 when the relation is linear and large y values are attached to small x values. Property 4 states that the value of r is unchanged when a

constant is added to each of the x variables (or to each of the y variables) or when each x variable (or each y variable) is multiplied by a positive constant. This property implies that r does not depend on the dimensions chosen to measure the data. For instance, the sample correlation coefficient between a person's height and weight does not depend on whether the height is measured in feet or in inches or whether the weight is measured in pounds or in kilograms. Also, if one of the values in the pair is temperature, then the sample correlation coefficient is the same whether it is measured in Fahrenheit or in Celsius.

The absolute value of the sample correlation coefficient r (that is, $|r|$, its value without regard to its sign) is a measure of the strength of the linear relationship between the x and the y values of a data pair. A value of $|r|$ equal to 1 means that there is a perfect linear relation — that is, a straight line can pass through all the data points (x_i, y_i), $i = 1, \ldots, n$. A value of $|r|$ of around .8 means that the linear relation is relatively strong; although there is no straight line that passes through all of the data points, there is one that is "close" to them all. A value for $|r|$ of around .3 means that the linear relation is relatively weak.

The sign of r gives the direction of the relation. It is positive when the linear relation is such that smaller y values tend to go with smaller x values and larger y values with larger x values (and so a straight line approximation points upward), and it is negative when larger y values tend to go with smaller x values and smaller y values with larger x values (and so a straight line approximation points downward). Figure 2.14 displays scatter diagrams for data sets with various values of r.

EXAMPLE 2.6a Find the sample correlation coefficient for the data presented in Table 2.8.

SOLUTION A computation gives the solution

$$r = .4189$$

thus indicating a relatively weak positive correlation between the daily temperature and the number of defective items produced that day. ■

EXAMPLE 2.6b The following data give the resting pulse rates (in beats per minute) and the years of schooling of 10 individuals. A scatter diagram of these data is presented in Figure 2.15. The sample correlation coefficient for these data is $r = -.7638$. This negative correlation indicates that for this data set a high pulse rate is strongly associated with a small number of years in school, and a low pulse rate with a large number of years in school. ■

Person	1	2	3	4	5	6	7	8	9	10
Years of School	12	16	13	18	19	12	18	19	12	14
Pulse Rate	73	67	74	63	73	84	60	62	76	71

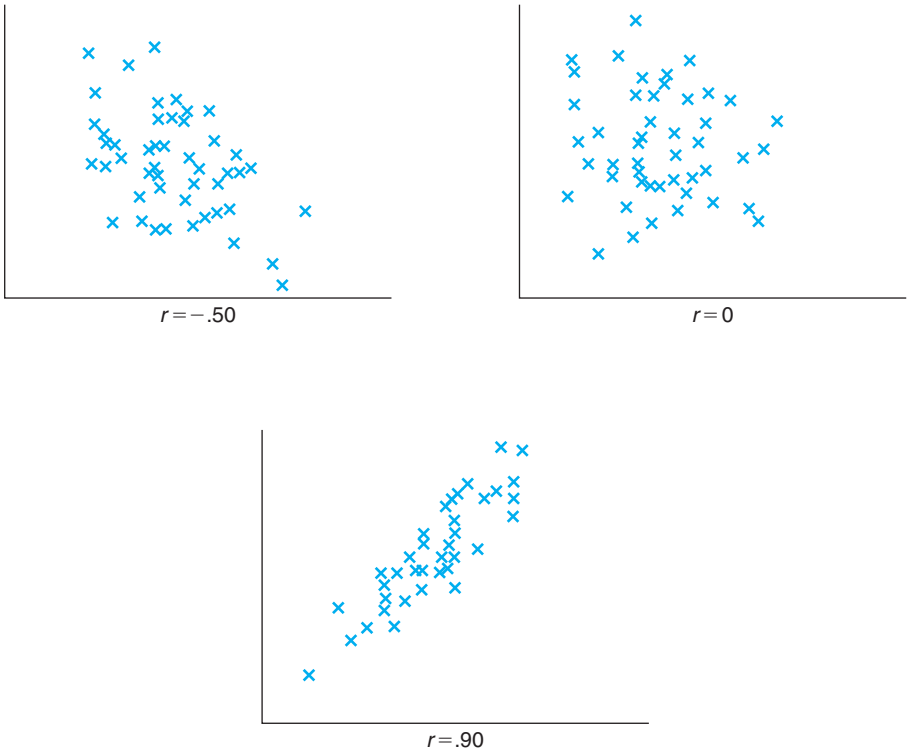

FIGURE 2.14 *Sample correlation coefficients.*

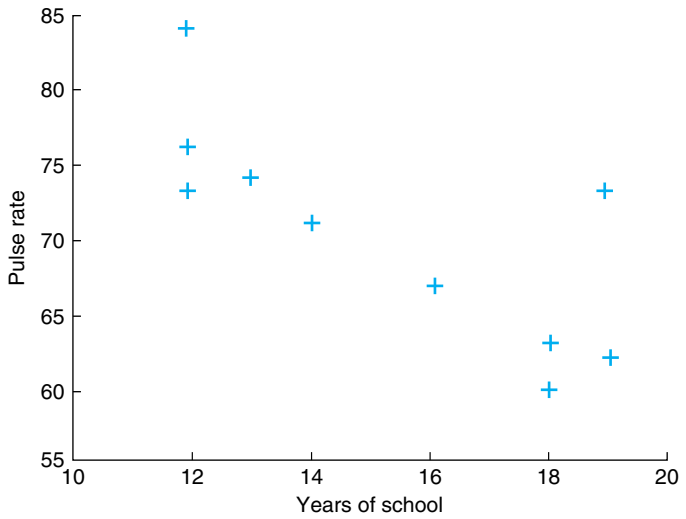

FIGURE 2.15 *Scatter diagram of years in school and pulse rate.*

Correlation Measures Association, Not Causation

The data set of Example 2.6b only considers 10 students and, as such, is not large enough for one to draw any firm conclusions about the relationship between years of school and pulse rate. Moreover, even if the data set were of larger size and with the same strong negative correlation between an individual's years of education and that individual's resting pulse rate, we would not be justified to conclude that additional years of school will directly reduce one's pulse rate. That is, whereas additional years of school tend to be associated with a lower resting pulse rate, this does not mean that it is a direct cause of it. Often, the explanation for such an association lies with an unexpressed factor that is related to both variables under consideration. In this instance, it may be that a person who has spent additional time in school is more aware of the latest findings in the area of health, and thus may be more aware of the importance of exercise and good nutrition; or it may be that it is not knowledge that is making the difference but rather it is that people who have had more education tend to end up in jobs that allow them more time for exercise and money for good nutrition. The strong negative correlation between years in school and resting pulse rate probably results from a combination of these as well as other underlying factors.

We will now prove the first three properties of the sample correlation coefficient r. That is, we will prove that $|r| \leq 1$ with equality when the data lie on a straight line. To begin, note that

$$\sum \left(\frac{x_i - \bar{x}}{s_x} - \frac{y_i - \bar{y}}{s_y} \right)^2 \geq 0 \tag{2.6.1}$$

or

$$\sum \frac{(x_i - \bar{x})^2}{s_x^2} + \sum \frac{(y_i - \bar{y})^2}{s_y^2} - 2 \sum \frac{(x_i - \bar{x})(y_i - \bar{y})}{s_x s_y} \geq 0$$

or

$$n - 1 + n - 1 - 2(n - 1)r \geq 0$$

showing that

$$r \leq 1$$

To see when $r = 1$, suppose first that the points (x_i, y_i), $i = 1, \ldots, n$ lie on the straight line

$$y_i = a + bx_i, \quad i = 1, \ldots, n$$

with positive slope b. If this is so, then

$$s_y^2 = b^2 s_x^2, \quad \bar{y} = a + b\bar{x}$$

showing that

$$b = \frac{s_y}{s_x}, \quad a = \bar{y} - \frac{s_y}{s_x}\bar{x}$$

Now, note also that $r = 1$ if and only if there is equality in Equation 2.6.1. That is, $r = 1$ if and only if for all i,

$$\frac{y_i - \bar{y}}{s_y} = \frac{x_i - \bar{x}}{s_x}$$

or, equivalently,

$$y_i = \bar{y} - \frac{s_y}{s_x}\bar{x} + \frac{s_y}{s_x}x_i$$

Consequently, $r = 1$ if and only if the data values (x_i, y_i) lie on a straight line having a positive slope.

To show that $r \geq -1$, with equality if and only if the data values (x_i, y_i) lie on a straight line having a negative slope, start with

$$\sum \left(\frac{x_i - \bar{x}}{s_x} + \frac{y_i - \bar{y}}{s_y} \right)^2 \geq 0$$

and use an argument analogous to the one just given.

Problems

1. The following is a sample of prices, rounded to the nearest cent, charged per gallon of standard unleaded gasoline in the San Francisco Bay area in June 1997.

3.88, 3.90, 3.93, 3.90, 3.93, 3.96, 3.88, 3.94, 3.96, 3.88, 3.94, 3.99, 3.98

Represent these data in

(a) a frequency table;
(b) a relative frequency line graph.

2. Explain how a pie chart can be constructed. If a data value had relative frequency r, at what angle would the lines defining its sector meet?

3. The following are the estimated oil reserves, in billions of barrels, for four regions in the Western Hemisphere.

United States	38.7
South America	22.6
Canada	8.8
Mexico	60.0

Represent these data in a pie chart.

4. The following table gives the average travel time to work for workers in each of the 50 states as well as the percentage of those workers who use public transportation.

(a) Represent the data relating to the average travel times in a histogram.
(b) Represent the data relating to the percentage of workers using public transportation in a stem and leaf plot.

Region, Division, and State	Means of Transportation to Work: Percent Using Public Transportation	Average Travel Time to Work[1] (minutes)
United States . .	**5.3**	**22.4**
Northeast	**12.8**	**24.5**
New England	5.1	21.5
Maine	0.9	19.0
New Hampshire .	0.7	21.9
Vermont	0.7	18.0
Massachusetts . .	8.3	22.7
Rhode Island . . .	2.5	19.2
Connecticut . . .	3.9	21.1
Middle Atlantic . . .	15.7	25.7
New York	24.8	28.6
New Jersey	8.8	25.3
Pennsylvania . . .	6.4	21.6
Midwest	**3.5**	**20.7**
East North Central . . .	4.3	21.7
Ohio	2.5	20.7
Indiana	1.3	20.4
Illinois	10.1	25.1
Michigan	1.6	21.2
Wisconsin	2.5	18.3
West North Central .	1.9	18.4
Minnesota	3.6	19.1
Iowa	1.2	16.2
Missouri	2.0	21.6

(continued)

Region, Division, and State	Means of Transportation to Work: Percent Using Public Transportation	Average Travel Time to Work[1] (minutes)
North Dakota	0.6	13.0
South Dakota	0.3	13.8
Nebraska	1.2	15.8
Kansas	0.6	17.2
South	**2.6**	**22.0**
South Atlantic	3.4	22.5
Delaware	2.4	20.0
Maryland	8.1	27.0
Virginia	4.0	24.0
West Virginia	1.1	21.0
North Carolina	1.0	19.8
South Carolina	1.1	20.5
Georgia	2.8	22.7
Florida	2.0	21.8
East South Central	1.2	21.1
Kentucky	1.6	20.7
Tennessee	1.3	21.5
Alabama	0.8	21.2
Mississippi	0.8	20.6
West South Central	2.0	21.6
Arkansas	0.5	19.0
Louisiana	3.0	22.3
Oklahoma	0.6	19.3
Texas	2.2	22.2
West	**4.1**	**22.7**
Mountain	2.1	19.7
Montana	0.6	14.8
Idaho	1.9	17.3
Wyoming	1.4	15.4
Colorado	2.9	20.7
New Mexico	1.0	19.1
Arizona	2.1	21.6
Utah	2.3	18.9
Nevada	2.7	19.8
Pacific	4.8	23.8
Washington	4.5	22.0
Oregon	3.4	19.6
California	4.9	24.6
Alaska	2.4	16.7
Hawaii	7.4	23.8

[1] *Excludes persons who worked at home.*

Source: U.S. Bureau of the Census. Census of Population and Housing, 1990.

5. Choose a book or article and count the number of words in each of the first 100 sentences. Present the data in a stem and leaf plot. Now choose another book or article, by a different author, and do the same. Do the two stem and leaf plots look similar? Do you think this could be a viable method for telling whether different articles were written by different authors?

6. The following table gives the number of commercial airline accidents and the total number of resulting fatalities in the United States in the years from 1985 to 2006.

   (a) Represent the number of yearly airline accidents in a frequency table.
   (b) Give a frequency polygon graph of the number of yearly airline accidents.
   (c) Give a cumulative relative frequency plot of the number of yearly airline accidents.
   (d) Find the sample mean of the number of yearly airline accidents.
   (e) Find the sample median of the number of yearly airline accidents.
   (f) Find the sample mode of the number of yearly airline accidents.
   (g) Find the sample standard deviation of the number of yearly airline accidents.

U.S. Airline Safety, Scheduled Commercial Carriers, 1985–2006

Year	Departures (millions)	Accidents	Fatalities	Year	Departures (millions)	Accidents	Fatalities
1985	6.1	4	197	1996	7.9	3	342
1986	6.4	2	5	1997	9.9	3	3
1987	6.6	4	231	1998	10.5	1	1
1988	6.7	3	285	1999	10.9	2	12
1989	6.6	11	278	2000	11.1	2	89
1990	7.8	6	39	2001	10.6	6	531
1991	7.5	4	62	2002	10.3	0	0
1992	7.5	4	33	2003	10.2	2	22
1993	7.7	1	1	2004	10.8	1	13
1994	7.8	4	239	2005	10.9	3	22
1995	8.1	2	166	2006	11.2	2	50

Source: National Transportation Safety Board.

7. (Use the table from Problem 6.)

   (a) Represent the number of yearly airline fatalities in a histogram.
   (b) Represent the number of yearly airline fatalities in a stem and leaf plot.
   (c) Find the sample mean of the number of yearly airline fatalities.
   (d) Find the sample median of the number of yearly airline fatalities.
   (e) Find the sample standard deviation of the number of yearly airline fatalities.

8. The sample mean of the weights of the adult women of town A is larger than the sample mean of the weights of the adult women of town B. Moreover, the sample mean of the weights of the adult men of town A is larger than the sample mean

of the weights of the adult men of town B. Can we conclude that the sample mean of the weights of the adults of town A is larger than the sample mean of the weights of the adults of town B? Explain your answer.

9. Using the table given in Problem 4, find the sample mean and sample median of the average travel time for those states in the

(a) northeast;
(b) midwest;
(c) south;
(d) west.

10. A total of 100 people work at company A, whereas a total of 110 work at company B. Suppose the total employee payroll is larger at company A than at company B.

(a) What does this imply about the median of the salaries at company A with regard to the median of the salaries at company B?
(b) What does this imply about the average of the salaries at company A with regard to the average of the salaries at company B?

11. The sample mean of the initial 99 values of a data set consisting of 198 values is equal to 120, whereas the sample mean of the final 99 values is equal to 100. What can you conclude about the sample mean of the entire data set

(a) Repeat when "sample mean" is replaced by "sample median."
(b) Repeat when "sample mean" is replaced by "sample mode."

12. The following table gives the number of pedestrians, classified according to age group and sex, killed in fatal road accidents in England in 1922.

(a) Approximate the sample means of the ages of the males.
(b) Approximate the sample means of the ages of the females.
(c) Approximate the quartiles of the males killed.
(d) Approximate the quartiles of the females killed.

Age	Number of Males	Number of Females
0–5	120	67
5–10	184	120
10–15	44	22
15–20	24	15
20–30	23	25
30–40	50	22
40–50	60	40
50–60	102	76
60–70	167	104
70–80	150	90
80–100	49	27

13. The following are the percentages of ash content in 12 samples of coal found in close proximity:

9.2, 14.1, 9.8, 12.4, 16.0, 12.6, 22.7, 18.9, 21.0, 14.5, 20.4, 16.9

Find the

(a) sample mean, and
(b) sample standard deviation of these percentages.

14. The sample mean and sample variance of five data values are, respectively, $\bar{x} = 104$ and $s^2 = 16$. If three of the data values are 102, 100, 105, what are the other two data values?

15. Suppose you are given the average pay of all working people in each of the 50 states of the United States.

(a) Do you think that the sample mean of the averages for the 50 states will equal the value given for the entire United States?
(b) If the answer to part (a) is no, explain what other information aside from just the 50 averages would be needed to determine the sample mean salary for the entire country. Also, explain how you would use the additional information to compute this quantity.

16. The following data represent the lifetimes (in hours) of a sample of 40 transistors:

112, 121, 126, 108, 141, 104, 136, 134
121, 118, 143, 116, 108, 122, 127, 140
113, 117, 126, 130, 134, 120, 131, 133
118, 125, 151, 147, 137, 140, 132, 119
110, 124, 132, 152, 135, 130, 136, 128

(a) Determine the sample mean, median, and mode.
(b) Give a cumulative relative frequency plot of these data.

17. An experiment measuring the percent shrinkage on drying of 50 clay specimens produced the following data:

18.2	21.2	23.1	18.5	15.6
20.8	19.4	15.4	21.2	13.4
16.4	18.7	18.2	19.6	14.3
16.6	24.0	17.6	17.8	20.2
17.4	23.6	17.5	20.3	16.6
19.3	18.5	19.3	21.2	13.9
20.5	19.0	17.6	22.3	18.4
21.2	20.4	21.4	20.3	20.1
19.6	20.6	14.8	19.7	20.5
18.0	20.8	15.8	23.1	17.0

(a) Draw a stem and leaf plot of these data.
(b) Compute the sample mean, median, and mode.
(c) Compute the sample variance.
(d) Group the data into class intervals of size 1 percent starting with the value 13.0, and draw the resulting histogram.
(e) For the grouped data acting as if each of the data points in an interval was actually located at the midpoint of that interval, compute the sample mean and sample variance and compare this with the results obtained in parts (b) and (c). Why do they differ?

18. A computationally efficient way to compute the sample mean and sample variance of the data set $x_1, x_2, \ldots, x_n$ is as follows. Let

$$\bar{x}_j = \frac{\sum_{i=1}^{j} x_i}{j}, \quad j = 1, \ldots, n$$

be the sample mean of the first j data values, and let

$$s_j^2 = \frac{\sum_{i=1}^{j} (x_i - \bar{x}_j)^2}{j-1}, \qquad j = 2, \ldots, n$$

be the sample variance of the first $j, j \geq 2$, values. Then, with $s_1^2 = 0$, it can be shown that

$$\bar{x}_{j+1} = \bar{x}_j + \frac{x_{j+1} - \bar{x}_j}{j+1}$$

and

$$s_{j+1}^2 = \left(1 - \frac{1}{j}\right) s_j^2 + (j+1)(\bar{x}_{j+1} - \bar{x}_j)^2$$

(a) Use the preceding formulas to compute the sample mean and sample variance of the data values 3, 4, 7, 2, 9, 6.
(b) Verify your results in part (a) by computing as usual.
(c) Verify the formula given above for $\bar{x}_{j+1}$ in terms of $\bar{x}_j$.

19. Use the data of Table 2.5 to find the

(a) 90 percentile of the average temperature for January;
(b) 75 percentile of the average temperature for July.

20. Find the quartiles of the following ages at death as given in obituaries of the New York Times in the 2 weeks preceding 1 August 2013.

92, 90, 92, 74, 69, 80, 94, 98, 65, 96, 84, 69, 86, 91, 88

74, 97, 85, 88, 68, 77, 94, 88, 65, 76, 75, 60

69, 97, 92, 85, 70, 80, 93, 91, 68, 82, 78, 89

21. The universities having the largest number of months in which they ranked in the top 10 for the number of google searches over the past 114 months (as of June 2013) are as follows.

University	Number of Months in Top 10
Harvard University	114
University of Texas, Austin	114
University of Michigan	114
Stanford University	113
University of California Los Angeles (UCLA)	111
University of California Berkeley	97
Penn State University	94
Massachusetts Institute of Technology (MIT)	66
University of Southern California (USC)	63
Ohio State University	52
Yale University	48
University of Washington	33

(a) Find the sample mean of the data.
(b) Find the sample variance of the data.
(c) Find the sample quartiles of the data.

22. Use the part of the table given in Problem 4 that gives the percentage of workers in each state that use public transportation to get to work to draw a box plot of these 50 percentages.

23. Represent the data of Problem 20 in a box plot.

24. The average particulate concentration, in micrograms per cubic meter, was measured in a petrochemical complex at 36 randomly chosen times, with the following concentrations resulting:

5, 18, 15, 7, 23, 220, 130, 85, 103, 25, 80, 7, 24, 6, 13, 65, 37, 25,

24, 65, 82, 95, 77, 15, 70, 110, 44, 28, 33, 81, 29, 14, 45, 92, 17, 53

(a) Represent the data in a histogram.
(b) Is the histogram approximately normal?

25. A chemical engineer desiring to study the evaporation rate of water from brine evaporation beds obtained data on the number of inches of evaporation in each of 55 July days spread over 4 years. The data are given in the following stem and leaf plot, which shows that the smallest data value was .02 inches, and the largest .56 inches.

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

Find the

(a) sample mean;
(b) sample median;
(c) sample standard deviation of these data.
(d) Do the data appear to be approximately normal?
(e) What percentage of data values are within 1 standard deviation of the mean?

26. The following are the grade point averages of 30 students recently admitted to the graduate program in the Department of Industrial Engineering and Operations Research at the University of California at Berkeley.

3.46, 3.72, 3.95, 3.55, 3.62, 3.80, 3.86, 3.71, 3.56, 3.49, 3.96, 3.90, 3.70, 3.61, 3.72, 3.65, 3.48, 3.87, 3.82, 3.91, 3.69, 3.67, 3.72, 3.66, 3.79, 3.75, 3.93, 3.74, 3.50, 3.83

(a) Represent the preceding data in a stem and leaf plot.
(b) Calculate the sample mean $\bar{x}$.
(c) Calculate the sample standard deviation s.
(d) Determine the proportion of the data values that lies within $\bar{x} \pm 1.5s$ and compare with the lower bound given by Chebyshev's inequality.
(e) Determine the proportion of the data values that lies within $\bar{x} \pm 2s$ and compare with the lower bound given by Chebyshev's inequality.

27. Do the data in Problem 26 appear to be approximately normal? For parts (c) and (d) of this problem, compare the approximate proportions given by the empirical rule with the actual proportions.

28. Would you expect that a histogram of the weights of all the members of a health club would be approximately normal?

29. Use the data of Problem 16.

(a) Compute the sample mean and sample median.
(b) Are the data approximately normal?
(c) Compute the sample standard deviation s.
(d) What percentage of the data fall within $\bar{x} \pm 1.5s$?
(e) Compare your answer in part (d) to that given by the empirical rule.
(f) Compare your answer in part (d) to the bound given by Chebyshev's inequality.

30. The following are the heights and starting salaries of 12 law school classmates whose law school examination scores were roughly the same.

Height	Salary
64	91
65	94
66	88
67	103
69	77
70	96
72	105
72	88
74	122
74	102
75	90
76	114

(a) Represent these data in a scatter diagram.
(b) Find the sample correlation coefficient.

31. A random sample of individuals were rated as to their standing posture. In addition, the numbers of days of back pain each had experienced during the past year were also recorded. Surprisingly to the researcher these data indicated a positive correlation between good posture and number of days of back pain. Does this indicate that good posture causes back pain?

32. If for each of the fifty states we plot the paired data consisting of the average income of residents of the state and the number of foreign-born immigrants who reside in the state, then the data pairs will have a positive correlation. Can we conclude that immigrants tend to have higher incomes than native-born Americans? If not, how else could this phenomenon be explained?

33. A random group of 12 high school juniors were asked to estimate the average number of hours they study each week. The following give these hours along with the student's grade point average.

Hours	GPA
6	2.8
14	3.2
3	3.1
22	3.6
9	3.0
11	3.3
12	3.4
5	2.7
18	3.1
24	3.8
15	3.0
17	3.9

Find the sample correlation coefficient between hours reported and GPA.

34. Verify property **3** of the sample correlation coefficient.

35. Verify property **4** of the sample correlation coefficient.

36. In a study of children in grades 2 through 4, a researcher gave each student a reading test. When looking at the resulting data the researcher noted a positive correlation between a student's reading test score and height. The researcher concluded that taller children read better because they can more easily see the blackboard. What do you think?

37. A recent study yielded a positive correlation between breast-fed babies and scores on a vocabulary test taken at age 6. Discuss the potential difficulties in interpreting the results of this study.