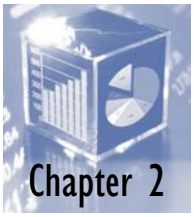

DESCRIPTIVE STATISTICS

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 a lower 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 $47,000 was received by four of the graduates, whereas the highest salary of $60,000 was received by a single student. The most common starting salary was $52,000, and was received by 10 of the students.

TABLE 2.1 *Starting Yearly Salaries*

Starting Salary	Frequency
47	4
48	1
49	3
50	5
51	8
52	10
53	0
54	5
56	2
57	3
60	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

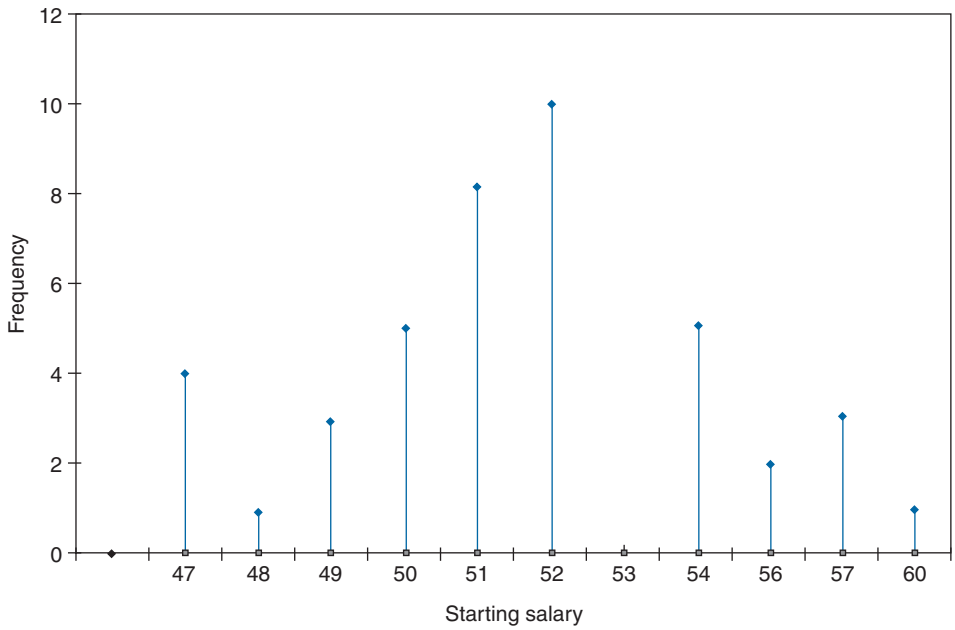

FIGURE 2.1 *Starting salary data.*

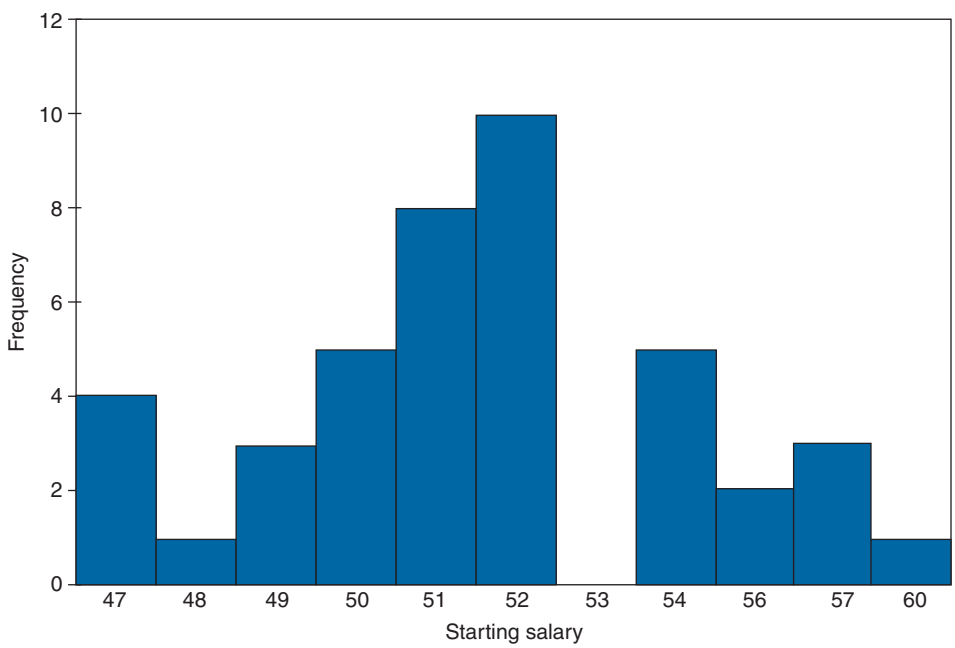

FIGURE 2.2 *Bar graph for starting salary data.*

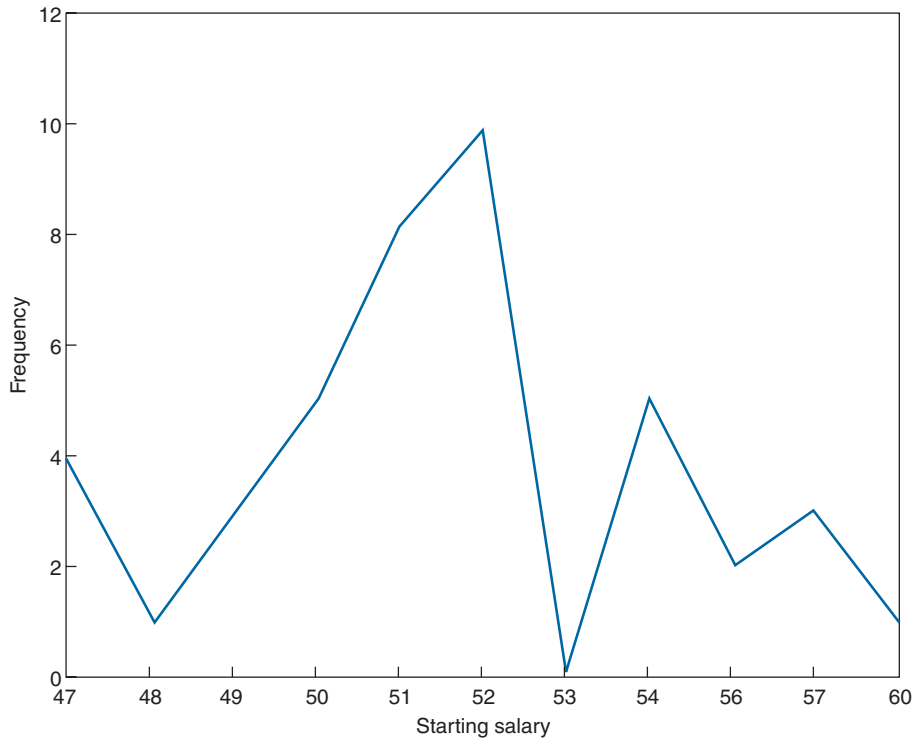

FIGURE 2.3 *Frequency polygon for starting salary data.*

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

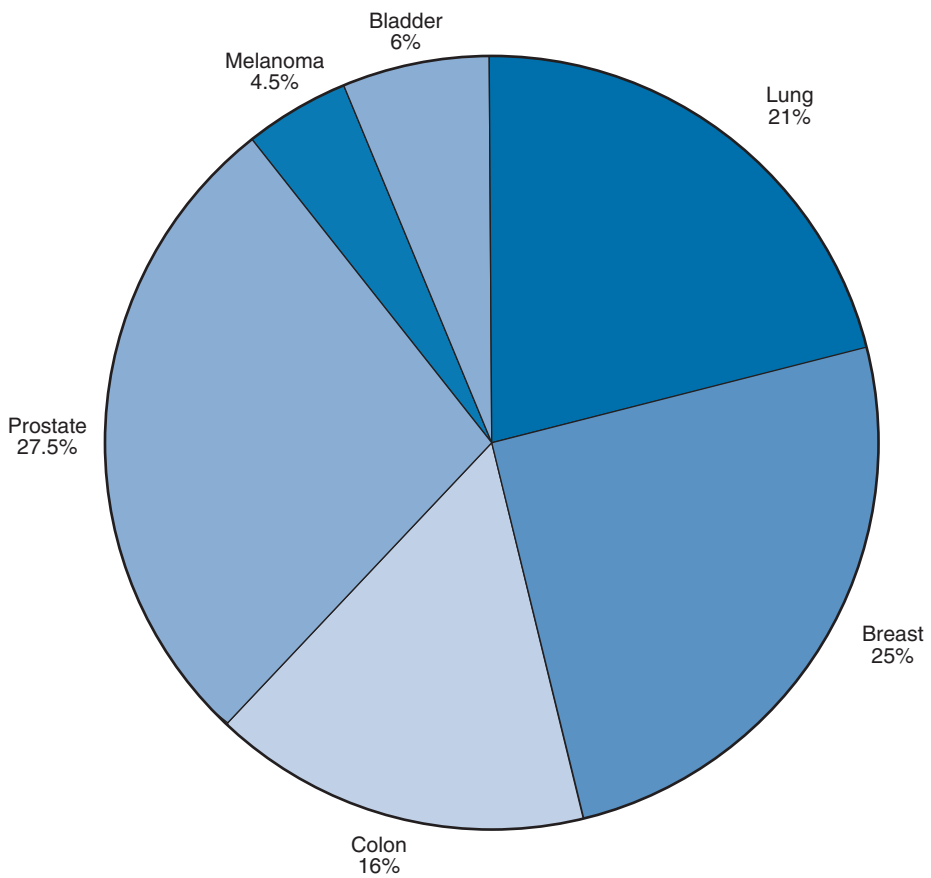

FIGURE 2.4

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

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

frequencies of each class being too small for a pattern to be discernible. Although 5 to 10 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

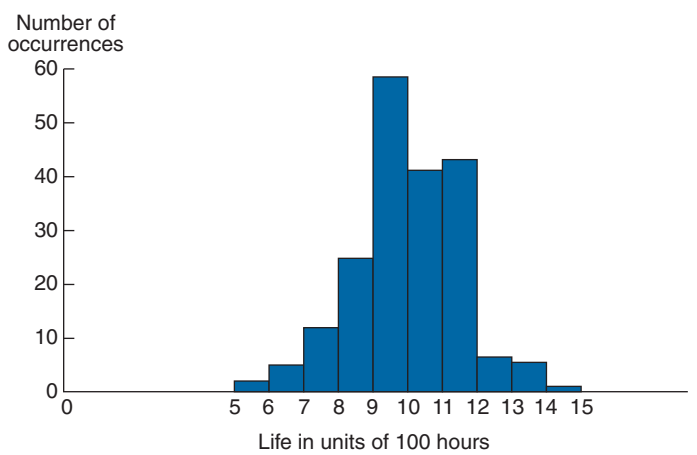

FIGURE 2.5 *A frequency histogram.*

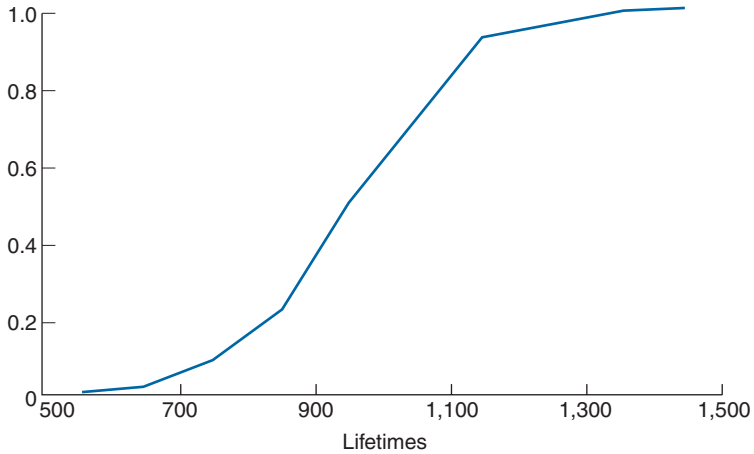

FIGURE 2.6 *A cumulative frequency plot.*

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 1982 to 1991 were as follows:

$$284,\ 280,\ 277,\ 282,\ 279,\ 285,\ 281,\ 283,\ 278,\ 277$$

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 $y_i = x_i - 280$:

$$4, 0, -3, 2, -1, 5, 1, 3, -2, -3$$

Because the arithmetic average of the transformed data set is

$$\bar{y} = 6/10$$

it follows that

$$\bar{x} = \bar{y} + 280 = 280.6 \quad \blacksquare$$

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.

A: 3, 4, 6, 7, 10 **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 1985 to 1993.

Year	1985	1986	1987	1988	1989	1990	1991	1992	1993
Accidents	22	22	26	28	27	25	30	29	24

Source: Civil Aviation Statistics of the World, annual.

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

SOLUTION Let us start by subtracting 22 from each value, to obtain the new data set:

$$0, 0, 4, 6, 5, 3, 8, 7, 2$$

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

$$\sum_{i=1}^{n} y_i = 35, \qquad \sum_{i=1}^{n} y_i^2 = 16 + 36 + 25 + 9 + 64 + 49 + 4 = 203$$

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

$$s^2 = \frac{203 - 9(35/9)^2}{8} \approx 8.361 \quad \blacksquare$$

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 $100p$ percent of the data are less than or equal to it and $100(1-p)$ 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.

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 520,947.

(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

$$\frac{1{,}151{,}977 + 1{,}524{,}249}{2} = 1{,}338{,}113 \quad \blacksquare$$

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*.

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,

TABLE 2.6 *Population of 25 Largest U.S. Cities, 1994*

Rank	City	Population
1	New York, NY	7,333,253
2	Los Angeles, CA	3,448,613
3	Chicago, IL	2,731,743
4	Houston, TX	1,702,086
5	Philadelphia, PA	1,524,249
6	San Diego, CA	1,151,977
7	Phoenix, AR	1,048,949
8	Dallas, TX	1,022,830
9	San Antonio, TX	998,905
10	Detroit, MI	992,038
11	San Jose, CA	816,884
12	Indianapolis, IN	752,279
13	San Francisco, CA	734,676
14	Baltimore, MD	702,979
15	Jacksonville, FL	665,070
16	Columbus, OH	635,913
17	Milwaukee, WI	617,044
18	Memphis, TN	614,289
19	El Paso, TX	579,307
20	Washington, D.C.	567,094
21	Boston, MA	547,725
22	Seattle, WA	520,947
23	Austin, TX	514,013
24	Nashville, TN	504,505
25	Denver, CO	493,559

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.

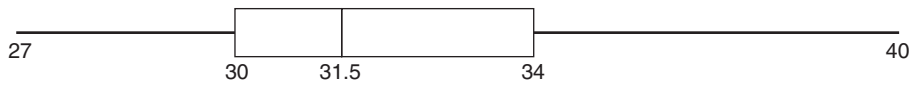

FIGURE 2.7 *A box plot.*

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

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 27 to a high value of 40. The value of the first quartile (equal to the value of the 11th smallest on the list) is 30; the value of the second quartile (equal to the average of the 21st and 22nd smallest values) is 31.5; and the value of the third quartile (equal to the value of the 32nd smallest on the list) is 34. 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 $N(S_k)$ be the number of elements in the set S_k. Then, for any $k \ge 1$,

$$\frac{N(S_k)}{n} \ge 1 - \frac{n-1}{nk^2} > 1 - \frac{1}{k_2}$$

Proof

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

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

$$\frac{n-1}{nk^2} \ge 1 - \frac{N(S_k)}{n}$$

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 1999. A simple calculation gives that the sample mean and sample standard deviation of

TABLE 2.7 *Top 10 Selling Cars for 1999*

	1999	
1.	Toyota Camry	448,162
2.	Honda Accord	404,192
3.	Ford Taurus	368,327
4.	Honda Civic	318,308
5.	Chevy Cavalier	272,122
6.	Ford Escort	260,486
7.	Toyota Corolla	249,128
8.	Pontiac Grand Am	234,936
9.	Chevy Malibu	218,540
10.	Saturn S series	207,977

these data are

$$\bar{x} = 298{,}217.8, \qquad s = 124{,}542.9$$

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

$$\left(\bar{x} - \frac{3}{2}s,\ \bar{x} + \frac{3}{2}s\right) = (173{,}674.9,\ 422,760.67)$$

whereas, in actuality, 90 percent of the data falls within those 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

For $k > 0$,

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

T

U

V

W