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 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.

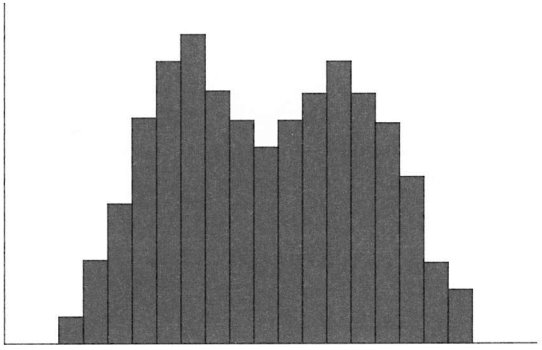

FIGURE 2.12 *Histogram of a bimodal data set.*

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

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.

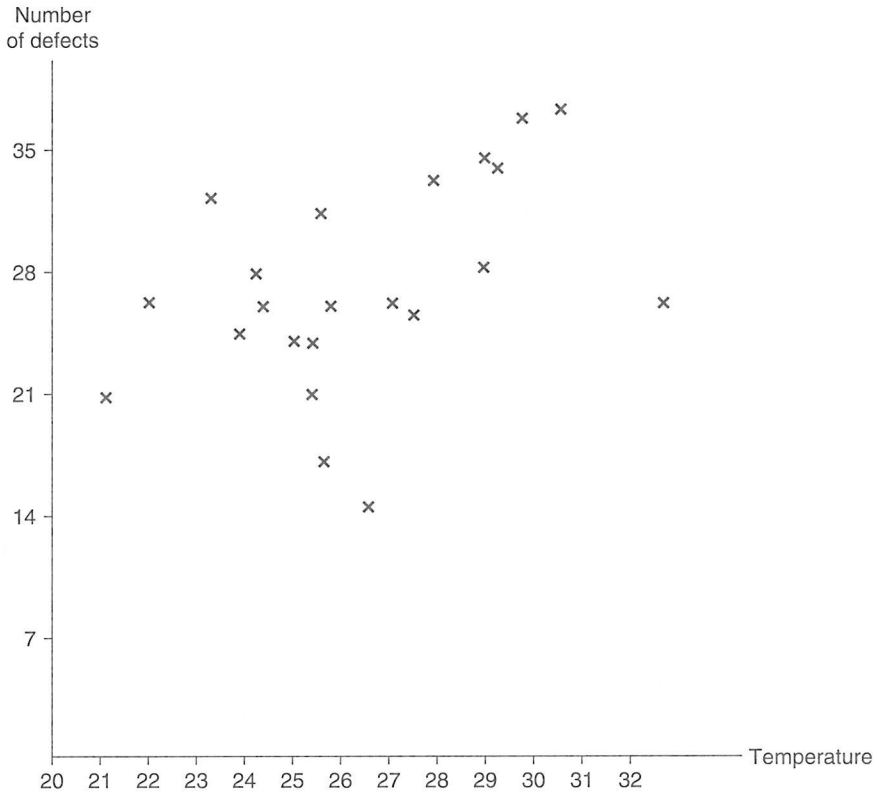

FIGURE 2.13 *A scatter diagram.*

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.

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.

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

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. $$-1 \leq r \leq 1$$
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.

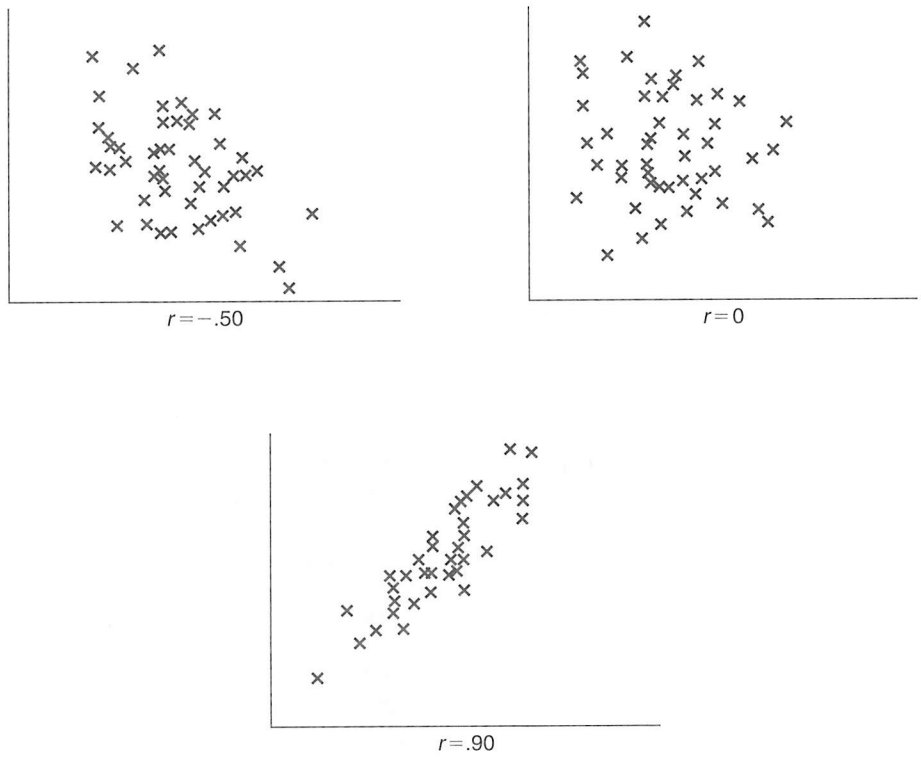

FIGURE 2.14 *Sample correlation coefficients.*

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. ■

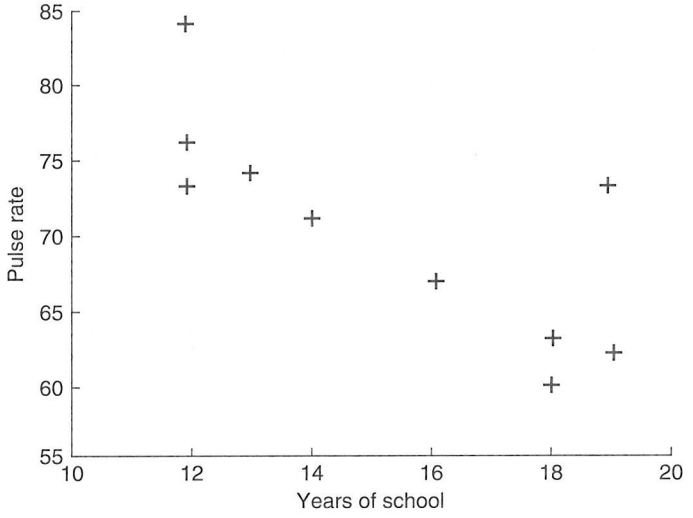

FIGURE 2.15 *Scatter diagram of years in school and pulse rate.*

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

Correlation Measures Association, Not Causation

The results of Example 2.6b indicate a strong negative correlation between an individual's years of education and that individual's resting pulse rate. However, this does not imply 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

$$r \leq 1$$

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,

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

That is, $r = 1$ 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.

(e) Find the sample standard deviation of the 1993 averages of the last 10 states listed.

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 Problem 8 to find the
(a) 10 percentile of the winning scores;
(b) 90 percentile of the winning scores.

20. Use the following table to find the quartiles of the average annual pay in the specified areas.

Average Annual Pay by New York State Metropolitan Areas, 1999

Rank	Amt.	Rank	Amt.
Albany-Sch'dy-Troy	$31,901	Nassau-Suffolk	$36,944
Binghamton	29,167	New York City	52,351
Buffalo-Niagara Falls	30,487	Newburgh, NY-PA	27,671
Dutchess County	35,256	Rochester	32,588
Elmira	26,603	Syracuse	30,423
Glens Falls	26,140	Utica-Rome	25,881
Jamestown	24,813	**US metro area avg.**	**$34,868**

Source: U.S. Bureau of Labor Statistics data.

21. Use the following figure (see next page), which gives the amounts of federal research money given to 15 universities in 1992, to answer this problem.
(a) Which universities were given more than $225 million?
(b) Approximate the sample mean of the amounts given to these universities.
(c) Approximate the sample variance of the amounts given to these universities.
(d) Approximate the quartiles of the amounts given to these universities.

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. Use the data concerning the first 10 states listed in the table given in Problem 15.
(a) Draw a scatter diagram relating the 1992 and 1993 salaries.
(b) Determine 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. Using data on the first 10 cities listed in Table 2.5, draw a scatter diagram and find the sample correlation coefficient between the January and July temperatures.

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 (reported in the May 5, 2008, *LA Times*) 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.

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.