CHAPTER 3

Using Statistics to Summarize Data Sets

I do hate averages. There is no greater mistake than to call arithmetic an exact science. There are permutations and aberrations discernible to minds entirely noble like mine; subtle variations which ordinary accountants fail to discover, hidden laws of numbers which it requires a mind like mine to perceive. For instance if you average numbers from the bottom up and then again from the top down, the result is always different.

A letter to the Mathematical Gazette (a 19th-century British mathematical journal)

The way to make sense out of raw data is to compare and contrast, to understand differences.

Gregory Bateson (in Steps to an Ecology of the Mind)

CONTENTS

Introductory Statistics, DOI: 10.1016/B978-0-12-374388-6.00003-X

Our objective in this chapter is to develop measures that can be used to summarize a data set. These measures, formally called *statistics*, are quantities whose values are determined by the data. We study the sample mean, sample median, and sample mode. These are all statistics that measure the center or middle value of a data set. Statistics that indicate the amount of variation in the data set are also considered. We learn about what it means for a data set to be normal, and we present an empirical rule concerning such sets. We also consider data sets consisting of paired values, and we present a statistic that measures the degree to which a scatter diagram of paired values can be approximated by a straight line.

3.1 INTRODUCTION

Modern-day experiments often track certain characteristics of thousands of individuals over time. For instance, in an attempt to learn about the health consequences of certain common practices, the medical statisticians R. Doll and A. B. Hill sent questionnaires in 1951 to all doctors in the United Kingdom and received 40,000 replies. Their questionnaire dealt with age, eating habits, exercise habits, and smoking habits. These doctors were then monitored for 10 years, and the causes of death of those who died were determined. As one can imagine, this study resulted in huge sets of data. For instance, even if we just focus on one component of the study at a single moment of time, such as the doctors' ages in 1951, the resulting data set of 40,000 values is vast. To obtain a feel for such a large data set, it is often necessary to summarize it by some suitably chosen measures. In this chapter, we introduce different statistics that can be used to summarize certain features of data sets.

To begin, suppose that we have in our possession sample data from some underlying population. Now, whereas in Chap. 2 we showed how to describe and portray data sets in their entirety, here we will be concerned with determining certain summary measures about the data. These summary measures are called *statistics*, where by a statistic we mean any numerical quantity whose value is determined by the data.

Definition *Numerical quantities computed from a data set are called* statistics.

We will be concerned with statistics that describe the central tendency of the data set; that is, they describe the center of the set of data values. Three different

statistics for describing this—the sample mean, sample median, and sample mode—will be presented in Secs. 3.2, 3.3, and 3.4, respectively. Once we have some idea of the center of a data set, the question naturally arises as to how much *variation* there is. That is, are most of the values close to the center, or do they vary widely about the center? In Sec. 3.5 we will discuss the sample variance and sample standard deviation, which are statistics designed to measure such variation.

In Sec. 3.6 we introduce the concept of a normal data set, which is a data set having a bell-shaped histogram. For data sets that are close to being normal, we present a rule that can be used to approximate the proportion of the data that is within a specified number of sample standard deviations from the sample mean.

In the first six sections of this chapter, we concern ourselves with data sets where each datum is a single value. However, in Sec. 3.7 we deal with paired data. That is, each data point will consist of an x value and a y value. For instance, the x value might represent the average number of cigarettes that an individual smoked per day, and the y value could be the age at which that individual died. We introduce a statistic called the *sample correlation coefficient* whose value indicates the degree to which data points having large x values also have large y values and correspondingly the degree to which those having small x values also have small y values.

The Doll–Hill study yielded the result that only about 1 in 1000 nonsmoking doctors died of lung cancer. For heavy smokers the figure was 1 in 8. In addition, death rates from heart attacks were 50 percent higher for smokers.

3.2 SAMPLE MEAN

Suppose we have a sample of n data points whose values we designate by $x_1, x_2, \ldots, x_n$. One statistic for indicating the center of this data set is the *sample mean*, defined to equal the arithmetic average of the data values.

Definition *The* sample mean, *which we designate by* $\bar{x}$ *(pronounced "x bar"), is defined by*

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

■ Example 3.1

The average fuel efficiencies, in miles per gallon, of cars sold in the United States in the years 1999 to 2003 were

$$28.2, 28.3, 28.4, 28.5, 29.0$$

Find the sample mean of this set of data.

Solution

The sample mean $\bar{x}$ is the average of the five data values. Thus,

$$\bar{x} = \frac{28.2 + 28.3 + 28.4 + 28.5 + 29.0}{5} = \frac{142.4}{5} = 28.48$$

Note from this example that whereas the sample mean is the average of all the data values, it need not itself be one of them. ■

Consider again the data set $x_1, x_2, \ldots, x_n$. If each data value is increased by a constant amount c, then this causes the sample mean also to be increased by c. Mathematically, we can express this by saying that if

$$y_i = x_i + c \quad i = 1, \ldots, n$$

then

$$\bar{y} = \bar{x} + c$$

where $\bar{y}$ and $\bar{x}$ are the sample means of the y_i and the x_i, respectively. Therefore, when it is convenient, we can compute $\bar{x}$ by first adding c to all the data values, then computing the sample mean $\bar{y}$ of the new data, and finally subtracting c from $\bar{y}$ to obtain $\bar{x}$. Since it is sometimes a lot easier to work with the transformed rather than the original data, this can greatly simplify the computation of $\bar{x}$. Our next example illustrates this point.

■ Example 3.2

The winning scores in the U.S. Masters Golf Tournament in the years from 1981 to 1990 were as follows:

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

Find the sample mean of these winning scores.

Solution

Rather than directly adding the preceding numbers, first we subtract 280 from (that is, add $c = -280$ to) each one to obtain the following transformed data:

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

The sample mean of these transformed data, call it $\bar{y}$, is

$$\bar{y} = \frac{0 + 4 + 0 - 3 + 2 - 1 + 5 + 1 + 3 - 2}{10} = \frac{9}{10}$$

Adding 280 to $\bar{y}$ shows that the sample mean of the original data is

$$\bar{x} = 280.9$$

■

If each data value is multiplied by c, then so is the sample mean. That is, if

$$y_i = cx_i \quad i = 1, \ldots, n$$

then

$$\bar{y} = c\bar{x}$$

For instance, suppose that the sample mean of the height of a collection of individuals is 5.0 feet. Suppose that we now want to change the unit of measurement from feet to inches. Then since each new data value is the old value multiplied by 12, it follows that the sample mean of the new data is $12 \cdot 5 = 60$. That is, the sample mean is 60 inches.

Our next example considers the computation of the sample mean when the data are arranged in a frequency table.

Example 3.3

The number of suits sold daily by a women's boutique for the past 6 days has been arranged in the following frequency table:

Value	Frequency
3	2
4	1
5	3

What is the sample mean?

Solution

Since the original data set consists of the 6 values

$$3, 3, 4, 5, 5, 5$$

it follows that the sample mean is

$$\begin{aligned} \bar{x} &= \frac{3+3+4+5+5+5}{6} \\ &= \frac{3 \times 2 + 4 \times 1 + 5 \times 3}{6} \\ &= \frac{25}{6} \end{aligned}$$

That is, the sample mean of the number of suits sold daily is 4.25. ■

In Example 3.3 we have seen that when the data are arranged in a frequency table, the sample mean can be expressed as the sum of the products of the distinct values and their frequencies, all divided by the size of the data set. This result holds in general. To see this, suppose the data are given in a frequency table that lists k distinct values $x_1, x_2, \ldots, x_k$ with respective frequencies $f_1, f_2, \ldots, f_k$. It follows that the data set consists of n observations, where $n = \sum_{i=1}^{k} f_i$ and where the value x_i appears f_i times for $i = 1, 2, \ldots, k$. Hence, the sample mean for this data set is

$$\begin{aligned}\bar{x} &= \frac{x_1 + \cdots + x_1 + x_2 + \cdots + x_2 + \cdots + x_k + \cdots + x_k}{n} \\ &= \frac{f_1 x_1 + f_2 x_2 + \cdots + f_k x_k}{n}\end{aligned} \tag{3.1}$$

Now, if $w_1, w_2, \ldots, w_k$ are nonnegative numbers that sum to 1, then

$$w_1 x_1 + w_2 x_2 + \cdots + w_k x_k$$

is said to be a *weighted average* of the values $x_1, x_2, \ldots, x_k$ with w_i being the weight of x_i. For instance, suppose that $k = 2$. Now, if $w_1 = w_2 = 1/2$, then the weighted average

$$w_1 x_1 + w_2 x_2 = \frac{1}{2} x_1 + \frac{1}{2} x_2$$

is just the ordinary average of x_1 and x_2. On the other hand, if $w_1 = 2/3$ and $w_2 = 1/3$, then the weighted average

$$w_1 x_1 + w_2 x_2 = \frac{2}{3} x_1 + \frac{1}{3} x_2$$

gives twice as much weight to x_1 as it does to x_2.

By writing Eq. (3.1) as

$$\bar{x} = \frac{f_1}{n} x_1 + \frac{f_2}{n} x_2 + \cdots + \frac{f_k}{n} x_k$$

we see that the sample mean $\bar{x}$ is a weighted average of the set of distinct values. The weight given to the value x_i is f_i/n, the proportion of the data values that is equal to x_i. Thus, for instance, in Example 3.3 we could have written that

$$\bar{x} = \frac{2}{6} \times 3 + \frac{1}{6} \times 4 + \frac{3}{6} \times 5 = \frac{25}{6}$$

■ Example 3.4

In a paper entitled "The Effects of Helmet Use on the Severity of Head Injuries in Motorcycle Accidents" (published in the *Journal of the American Statistical Association*, 1992, pp. 48–56), A. Weiss analyzed a sample of 770 similar motorcycle

accidents that occurred in the Los Angeles area in 1976 and 1977. Each accident was classified according to the severity of the head injury suffered by the motorcycle operator. The classification used was as follows:

Classification of accident	Interpretation
0	No head injury
1	Minor head injury
2	Moderate head injury
3	Severe, not life-threatening
4	Severe and life-threatening
5	Critical, survival uncertain at time of accident
6	Fatal

In 331 of the accidents the operator wore a helmet, whereas in the other 439 accidents the operator did not. The following are frequency tables giving the severities of the accidents that occurred when the operator was wearing and was not wearing a helmet.

Classification	Frequency of driver with helmet	Frequency of driver without helmet
0	248	227
1	58	135
2	11	33
3	3	14
4	2	3
5	8	21
6	1	6
	331	439

Find the sample mean of the head severity classifications for those operators who wore helmets and for those who did not.

Solution

The sample mean for those wearing helmets is

$$\bar{x} = \frac{0.248 + 1.58 + 2.11 + 3.3 + 4.2 + 5.8 + 6.1}{331} = \frac{143}{331} = 0.432$$

The sample mean for those who did not wear a helmet is

$$\bar{x} = \frac{0.227 + 1.135 + 2.33 + 3.14 + 4.3 + 5.21 + 6.6}{439} = \frac{396}{439} = 0.902$$

Therefore, the data indicate that those cyclists who were wearing a helmet suffered, on average, less severe head injuries than those who were not wearing a helmet. ■

3.2.1 Deviations

Again suppose that sample data consist of the n values $x_1, \ldots, x_n$ and that $\bar{x} = \sum_{i=1}^{n} x_i/n$ is the sample mean. The differences between each of the data values and the sample mean are called *deviations*.

Definition *The* deviations *are the differences between the data values and the sample mean. The value of the ith deviation is* $x_i - \bar{x}$.

A useful identity is that the sum of all the deviations must equal 0. That is,

$$\sum_{i=1}^{n} (x_i - \bar{x}) = 0$$

That this equality is true is seen by the following argument:

$$\begin{aligned} \sum_{i=1}^{n} (x_i - \bar{x}) &= \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \bar{x} \\ &= n\bar{x} - n\bar{x} \\ &= 0 \end{aligned}$$

This equality states that the sum of the positive deviations from the sample mean must exactly balance the sum of the negative deviations. In physical terms, this means that if n weights of equal mass are placed on a (weightless) rod at the points x_i, $i = 1, \ldots, n$, then $\bar{x}$ is the point at which the rod will be in balance. This balancing point is called the *center of gravity* (Fig. 3.1).

Historical Perspective

In the early days of sea voyages it was quite common for large portions of a ship's cargo to be either lost or damaged due to storms. To handle this potential loss, there was a standard agreement that all those having merchandise aboard the ship would contribute to pay for the value of all lost or damaged goods. The amount of money that each of them was called upon to pay was known as *havaria*, and from this Latin word derives our present word *average*. (Typically, if there were n shippers having damages $x_1, \ldots, x_n$, then the total loss was $x_1 + \cdots + x_n$ and the havaria for each was $(x_1 + \cdots + x_n)/n$.)

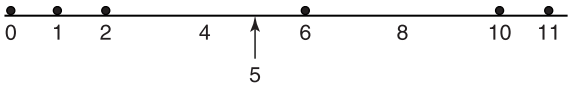


FIGURE 3.1
The center of gravity of 0, 1, 2, 6, 10, 11 is (0 + 1 + 2 + 6 + 10 + 11)/6 = 30/6 = 5.

■ Example 3.5

For the data of Example 3.1, the deviations from the sample mean of 28.48 are

$$x_1 - \bar{x} = 28.2 - 28.48 = -0.28$$
$$x_2 - \bar{x} = 28.3 - 28.48 = -0.18$$
$$x_3 - \bar{x} = 28.4 - 28.48 = -0.08$$
$$x_4 - \bar{x} = 28.5 - 28.48 = 0.02$$
$$x_5 - \bar{x} = 29.0 - 28.48 = 0.52$$

As a check, we note that the sum of the deviations is

$$-0.28 - 0.18 - 0.08 + 0.02 + 0.52 = 0$$ ■

PROBLEMS

1. The following data represent the scores on a statistics examination of a sample of students:

   87, 63, 91, 72, 80, 77, 93, 69, 75, 79, 70, 83, 94, 75, 88

   What is the sample mean?
2. The following data (from U.S. Department of Agriculture, *Food Consumption, Prices, and Expenditures*) give the U.S. per capita consumption (in pounds) of cheese in a sample of years.

Year	1965	1975	1985	1995	2001
Per capita consumption	10.0	14.8	23.4	26.4	30.1

   Find the sample mean of the given data.
3. The following data give the annual average number of inches of precipitation and the average number of days of precipitation in a sample of cities.

City	Average amount of precipitation	Average number of days
Albany, NY	35.74	134
Baltimore, MD	31.50	83
Casper, WY	11.43	95
Denver, CO	15.31	88
Fargo, ND	19.59	100
Houston, TX	44.76	105
Knoxville, TN	47.29	127
Los Angeles, CA	12.08	36
Miami, FL	57.55	129
New Orleans, LA	59.74	114
Pittsburgh, PA	36.30	154
San Antonio, TX	29.13	81
Wichita, KS	28.61	85

Source: National Oceanic and Atmospheric Administration.

(a) Find the sample mean of the average number of inches of precipitation.
(b) Find the sample mean of the average number of days of precipitation.

4. Consider five numbers. Suppose the mean of the first four numbers is 14.
(a) If the fifth number is 24, what is the mean of all five numbers?
(b) If the mean of all five numbers is 24, what is the fifth number?

5. 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.

6. Suppose that the sample mean of a set of 10 data points is $\bar{x} = 20$.
(a) If it is discovered that a data point having value 15 was incorrectly read as having value 13, what should be the revised value of $\bar{x}$?
(b) Suppose there is an additional data point whose value is 22. Will this increase or decrease the value of $\bar{x}$?
(c) Using the original data (and not the revised data in part (a)), what is the new value of $\bar{x}$ in part (b)?

7. The following table gives the number of cases of tetanus in the 27-country European community in the years from 1996 to 2004. Find the sample mean of these $27 \times 11 = 297$ data values.

Tetanus
Number of cases

	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005
EU-27	**352**	**309**	**290**	**288**	**263**	**220**	**188**	**219**	**177**	**98**
Belgium	3	1	0	1	1	3	1	1	2	3
Bulgaria	4	5	1	10	3	4	2	2	0	2
Czech Republic	0	2	1	0	1	3	0	0	0	0
Denmark	0	2	2	1	4	1	0	0	1	0
Germany	17	11	7	8	8	8	:	:	:	:
Estonia	1	2	1	1	1	1	0	1	0	0
Ireland	0	0	1	1	1	3	0	0	1	0
Greece	7	2	2	6	16	4	3	7	5	7
Spain	45	45	32	38	29	23	21	24	16	18
France	39	29	20	17	29	28	17	30	25	17
Italy	105	103	119	91	98	63	69	73	56	:
Cyprus	0	0	0	0	0	0	1	2	0	0
Latvia	4	3	1	3	2	1	0	0	1	0
Lithuania	2	3	2	6	0	1	1	4	1	4
Luxembourg	1	0	0	0	0	0	0	0	0	0
Hungary	11	12	12	20	10	8	5	4	1	3
Malta	1	1	1	0	1	1	2	1	1	0
Netherlands	2	2	3	1	5	0	:	4	:	:
Austria	0	0	0	0	0	0	:	0	:	:
Poland	46	37	22	21	14	21	20	30	25	15
Portugal	23	16	24	25	15	15	11	6	9	8
Romania	22	17	23	19	14	23	21	13	11	8
Slovenia	5	5	3	5	9	2	5	3	2	2
Slovakia	1	0	0	0	0	0	2	0	0	0
Finland	1	0	2	8	:	:	:	0	:	:
Sweden	3	3	2	2	0	1	0	0	0	1
United Kingdom	9	8	9	4	2	6	7	14	20	10

8. The following stem-and-leaf plot portrays the most recent 15 league bowling scores of the author of this text. Compute the sample mean.

18	2, 4, 7
17	0
16	1,9
15	2, 2, 4, 8, 8
14	
13	2, 1, 5, 5

9. Find the sample mean for this data set:

1, 2, 4, 7, 10, 12

Now find the sample means for the data sets

$$3, 6, 12, 21, 30, 36 \quad \text{and} \quad 6, 7, 9, 12, 15, 17$$

10. Suppose that $\bar{x}$ is the sample mean of the data set consisting of the data $x_1, \ldots, x_n$. If the data are transformed according to the formula

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

what is the sample mean of the data set $y_1, \ldots, y_n$? (In the equation, a and b are given constants.)

11. The following data give the total number of fires in Ontario, Canada, in the months of 2002:

$$6, 13, 5, 7, 7, 3, 7, 2, 5, 6, 9, 8$$

Find the sample mean of this data set.

12. The following data set specifies the total number of cars produced in the United States over a sample of years. The data are in units of 1000 cars. Find the sample mean of the number of cars sold annually in these years.

Year	1980	1985	1990	1995	2000	2002	2006
Number sold	8010	11,653	9783	11,985	12,832	12,326	11,264

Source: *Statistical Abstract of the United States, 2008.*

13. One-half the values of a sample are equal to 10, and the other half are equal to 20. What is the sample mean?

14. The following is a frequency table of the ages of a sample of members of a symphony for young adults.

Age value	Frequency
16	9
17	12
18	15
19	10
20	8

Find the sample mean of the given ages.

15. Half the values of a sample are equal to 10, one-sixth are equal to 20, and one-third are equal to 30. What is the sample mean?

16. There are two entrances to a parking lot. Student 1 counts the daily number of cars that pass through entrance 1, and student 2 does the same for entrance 2. Over 30 days, the data of student 1 yielded a

sample mean of 122, and the data of student 2 yielded a sample mean of 160. Over these 30 days, what was the daily average number of cars that entered the parking lot?

17. A company runs two manufacturing plants. A sample of 30 engineers at plant 1 yielded a sample mean salary of $33,600. A sample of 20 engineers at plant 2 yielded a sample mean salary of $42,400. What is the sample mean salary for all 50 engineers?
18. Suppose that we have two distinct samples of sizes n_1 and n_2. If the sample mean of the first sample is $\bar{x}_1$ and that of the second is $\bar{x}_2$, what is the sample mean of the combined sample of size $n_1 + n_2$?
19. Find the deviations for each of the three data sets of Prob. 9, and verify your answers by showing that in each case the sum of the deviations is 0.
20. Calculate the deviations for the data of Prob. 14 and check that they sum to 0.

3.3 SAMPLE MEDIAN

The following data represent the number of weeks after completion of a learn-to-drive course that it took a sample of seven people to obtain a driver's license:

$$2, 110, 5, 7, 6, 7, 3$$

The sample mean of this data set is $\bar{x} = 140/7 = 20$; and so six of the seven data values are quite a bit less than the sample mean, and the seventh is much greater. This points out a weakness of the sample mean as an indicator of the center of a data set—namely, its value is greatly affected by extreme data values.

A statistic that is also used to indicate the center of a data set but that is not affected by extreme values is the *sample median*, defined as the middle value when the data are ranked in order from smallest to largest. We will let m denote the sample median.

Definition *Order the data values from smallest to largest. If the number of data values is odd, then the* sample median *is the middle value in the ordered list; if it is even, then the* sample median *is the average of the two middle values.*

It follows from this definition that if there are three data values, then the sample median is the second-smallest value; and if there are four, then it is the average of the second- and the third-smallest values.

■ Example 3.6

The following data represent the number of weeks it took seven individuals to obtain their driver's licenses. Find the sample median.

$$2, 110, 5, 7, 6, 7, 3$$

Solution

First arrange the data in increasing order.

$$2, 3, 5, 6, 7, 7, 110$$

Since the sample size is 7, it follows that the sample median is the fourth-smallest value. That is, the sample median number of weeks it took to obtain a driver's license is $m = 6$ weeks. ■

■ Example 3.7

The following data represent the number of days it took 6 individuals to quit smoking after completing a course designed for this purpose.

$$1, 2, 3, 5, 8, 100$$

What is the sample median?

Solution

Since the sample size is 6, the sample median is the average of the two middle values; thus,

$$m = \frac{3 + 5}{2} = 4$$

That is, the sample median is 4 days. ■

In general, for a data set of n values, the sample median is the $[(n + 1)/2]$-smallest value when n is odd and is the average of the $(n/2)$-smallest value and the $(n/2 + 1)$-smallest value when n is even.

The sample mean and sample median are both useful statistics for describing the central tendency of a data set. The sample mean, being the arithmetic average, makes use of all the data values. The sample median, which makes use of only one or two middle values, is not affected by extreme values.

■ Example 3.8

The following data give the names of the National Basketball Association (NBA) individual scoring champions and their season scoring averages in each of the seasons from 1992 to 2008.

(a) Find the sample median of the scoring averages.
(b) Find the sample mean of the scoring averages.

1992–93	Michael Jordan, Chicago Bulls	32.6
1993–94	David Robinson, San Antonio Spurs	29.8
1994–95	Shaquille O'Neal, Orlando Magic	29.3
1995–96	Michael Jordan, Chicago Bulls	30.4
1996–97	Michael Jordan, Chicago Bulls	29.6
1997–98	Michael Jordan, Chicago Bulls	28.7
1998–99	Allen Iverson, Philadelphia 76ers	26.8
1999–00	Shaquille O'Neal, L.A. Lakers	29.7
2000–01	Allen Iverson, Philadelphia 76ers	31.1
2001–02	Allen Iverson, Philadelphia 76ers	31.4
2002–03	Tracy McGrady, Orlando Magic	32.1
2003–04	Tracy McGrady, Orlando Magic	28.0
2004–05	Allen Iverson, Philadelphia 76ers	30.7
2005–06	Kobe Bryant, L.A. Lakers	35.4
2006–07	Kobe Bryant, Los Angeles Lakers	31.6
2007–08	Lebron James, Cleveland Cavaliers	30.0
2008–09	Dwyane Wade, Miami Heat	30.2

Solution

(a) Since there are 17 data values, the sample median is the 9th smallest. Therefore, the sample median is

$$m = 30.2$$

(b) The sum of all 17 values is 517.4, and so the sample mean is

$$\bar{x} = \frac{517.4}{17} \approx 30.435$$

■

Historical Perspective

The Dutch mathematician Christian Huyghens was one of the early developers of the theory of probability. In 1669 his brother Ludwig, after studying the mortality tables of the time, wrote to his famous older brother that "I have just been making a table showing how long people have to live. ... Live well! According to my calculations you will live to be about $56\frac{1}{2}$ and I to 55." Christian, intrigued, also looked at the mortality tables but came up with different estimates for how long both he and his brother would live. Why? Because they were looking at different statistics. Ludwig was basing his estimates on the sample median while Christian was basing his on the sample mean!

For data sets that are roughly symmetric about their central values, the sample mean and sample median will have values close to each other. For instance, the data

$$4, 6, 8, 8, 9, 12, 15, 17, 19, 20, 22$$

are roughly symmetric about the value 12, which is the sample median. The sample mean is $\bar{x} = 140/11 = 12.73$, which is close to 12.

The question as to which of the two summarizing statistics is the more informative depends on what you are interested in learning from the data set. For instance, if a city government has a flat-rate income tax and is trying to figure out how much income it can expect, then it would be more interested in the sample mean of the income of its citizens than in the sample median (why is this?). On the other hand, if the city government were planning to construct some middle-income housing and were interested in the proportion of its citizens who would be able to afford such housing, then the sample median might be more informative (why is this?).

Although it is interesting to consider whether the sample mean or sample median is more informative in a particular situation, note that we need never restrict ourselves to a knowledge of just one of these quantities. They are both important, and thus both should always be computed when a data set is summarized.

PROBLEMS

1. The following are the total yardages of a sample of 12 municipal golf courses:

   7040, 6620, 6050, 6300, 7170, 5990, 6330, 6780, 6540, 6690, 6200, 6830

   (a) Find the sample median.
   (b) Find the sample mean.
2. (a) Determine the sample median of the data set

   14, 22, 8, 19, 15, 7, 8, 13, 20, 22, 24, 25, 11, 9, 14

   (b) Increase each value in (a) by 5, and find the new sample median.
   (c) Multiply each value in (a) by 3, and find the new sample median.
3. If the median of the data set $x_i, i = 1, \ldots, n$, is 10, what is the median of the data set $2x_i + 3, i = 1, \ldots, n$?
4. The following are the speeds of 40 cars as measured by a radar device on a city street:

22, 26, 31, 38, 27, 29, 33, 40, 36, 27, 25, 42, 28, 19, 28, 26, 33, 26, 37, 22, 31, 30, 44, 29, 25, 17, 46, 28, 31, 29, 40, 38, 26, 43, 45, 21, 29, 36, 33, 30

Find the sample median.

5. The following presents the male and female suicide rates per 100,000 population for a variety of countries.

Suicide Rates per 100,000 Population

Sex	United States	Australia	Austria	Canada	Denmark	France
Female	5.4	5.1	15.8	5.4	20.6	12.7
Male	19.7	18.2	42.1	20.5	35.1	33.1

Sex	Italy	Japan	Netherlands	Poland	Sweden	U.K.	W. Germany
Female	4.3	14.9	8.1	4.4	11.5	5.7	12.0
Male	11.0	27.8	14.6	22.0	25.0	12.1	26.6

Source: World Health Organization, *World Health Statistics*.

(a) Find the sample median of the male suicide rates.
(b) Find the sample median of the female suicide rates.
(c) Find the sample mean of the male suicide rates.
(d) Find the sample mean of the female suicide rates.

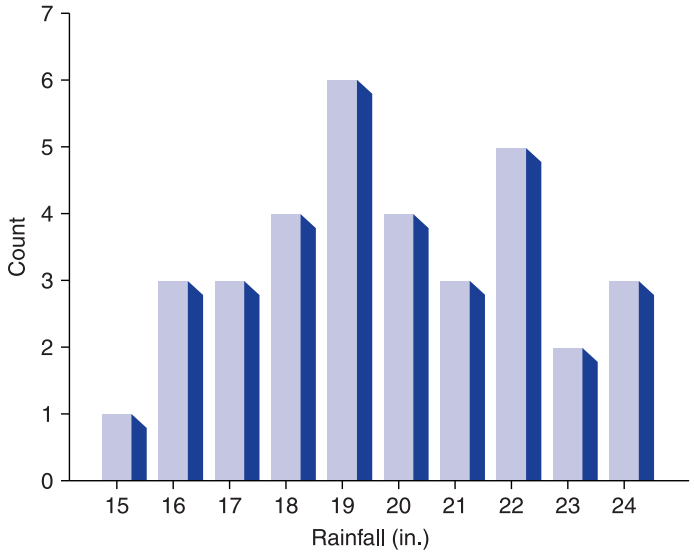
6. Find the sample median of the average annual number of days of precipitation in the cities noted in Prob. 3 of Sec. 3.2.
7. Find the sample median of the average annual number of inches of precipitation in the cities noted in Prob. 3 of Sec. 3.2.
8. Find the sample median of the data presented in Prob. 8 of Sec. 2.3.
9. Use the table on death rates preceding Prob. 9 of Sec. 2.3 to find the sample median of the death rates due to

(a) Falls
(b) Poisoning
(c) Drowning

10. The sample median of 10 distinct values is 5. What can you say about the new sample median if

(a) An additional datum whose value is 7 is added to the data set?
(b) Two additional data values—3 and 42—are added to the data set?

11. The histogram in the figure on the following page describes the annual rainfall, in inches, over the last 34 years in a certain western city. Since the raw data are not recoverable from a histogram, we cannot use them to exactly compute the value of the sample mean and sample median. Still, based on this histogram, what is the largest possible value of



(a) The sample mean?
(b) The sample median?
What is the smallest possible value of
(c) The sample mean?
(d) The sample median?
(e) The actual data follow:

15.2, 16.1, 16.5, 16.7, 17.2, 17.5, 17.7, 18.3, 18.6, 18.8, 18.9, 19.1,
19.2, 19.2, 19.6, 19.8, 19.9, 20.2, 20.3, 20.3, 20.8, 21.1, 21.4, 21.7,
22.2, 22.5, 22.5, 22.7, 22.9, 23.3, 23.6, 24.1, 24.5, 24.9

Determine the sample mean and sample median and see that they are consistent with your previous answers.

12. 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 sample mean of the salaries at company A with regards to the sample mean of the salaries at company B?
(b) What does this imply about the sample median of the salaries at company A with regards to the sample median of the salaries at company B?

13. Using the data from Example 3.4, compute the sample medians of the severity of head injuries suffered by motorcycle operators who were wearing and who were not wearing helmets.

14. In the following situations, which do you think is a more informative statistic, the sample mean or the sample median?

(a) In order to decide whether to discontinue a bus service from Rochester to New York City, an executive studies the number of riders on a sample of days.

(b) To determine how present-day college-bound students compare with those of earlier years, a sample of entrance examination scores from several years is consulted.

(c) A lawyer representing a defendant in a jury trial is studying the IQ scores of the jurors who were selected.

(d) You purchased your home 6 years ago in a small suburban community for $105,000, which was both the mean and the median price for all homes sold that year in that community. However, in the last couple of years some new, more expensive homes have been built. To get an idea of the present value of your home, you study recent sales prices of homes in your community.

15. Women make up the following percentages of the workforce in the 14 occupations listed.

Occupation	Percentage women	Occupation	Percentage women
Corporate executives	36.8	Doctors	17.6
Nurses	94.3	Lawyers	18.0
Sales supervisors	30.5	Elementary school teachers	85.2
Sales workers	68.6	Postal clerks	43.5
Firefighters	1.9	Police workers	10.9
Cleaning jobs	41.5	Construction supervisors	1.6
Construction workers	2.8	Truck drivers	2.1

For these percentages find

(a) The sample mean

(b) The sample median

It also turns out that women make up 44.4 percent of the total workforce for these occupations. Is this consistent with your answers in (a) and (b)? Explain!

16. Using data concerning the first 30 students in App. A, find the sample median and the sample mean for

(a) Weight

(b) Cholesterol

(c) Blood pressure

17. The following table gives the median age at first marriage in the years 1992 to 2002.

(a) Find the sample median of the men's median age.
(b) Find the sample median of the women's median age.

U.S. Median Age at First Marriage

Year	Men	Women	Year	Men	Women
2002	26.9	25.3	1996	27.1	24.8
2001	26.9	25.1	1995	26.9	24.5
2000	26.8	25.1	1994	26.7	24.5
1999	26.9	25.1	1993	26.5	24.5
1998	26.7	25.0	1992	26.5	24.4
1997	26.8	25.0			

3.3.1 Sample Percentiles

The sample median is a special type of statistic known as a *sample* $100p$ *percentile*, where p is any fraction between 0 and 1. Loosely speaking, a sample $100p$ percentile is the value such that $100p$ percent of the data values are less than it and $100(1 - p)$ percent of the values are greater than it.

Definition *The* sample $100p$ percentile *is that data value having the property that at least* $100p$ *percent of the data are less than or equal to it and at least* $100(1 - p)$ *percent of the data values 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 values.*

Note that the sample median is the sample 50th percentile. That is, it is the sample $100p$ percentile when $p = 0.50$.

Suppose the data from a sample of size n are arranged in increasing order from smallest to largest. To determine the sample $100p$ percentile, we must determine the data value such that

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

Now if np is not an integer, then the only data value satisfying these requirements is the one whose position is the smallest integer greater than np. For instance, suppose we want the sample 90th percentile from a sample of size $n = 12$. Since $p = 0.9$, we have $np = 10.8$ and $n(1 - p) = 1.2$. Thus, we require those data values for which

1. At least 10.8 values are less than or equal to it (and so the data value must be in position 11 or higher).
2. At least 1.2 values are greater than or equal to it (and so it must be in position 11 or lower).

Clearly, the only data value that satisfies both requirements is the one that is in position 11, and thus this is the sample 90th percentile.

On the other hand, if np is an integer, then both the value in position np and the value in position $np + 1$ satisfy requirements (1) and (2); and so the sample $100p$ percentile value would be the average of these two data values. For instance, suppose we wanted the sample 95th percentile from a data set of $n = 20$ values. Then both the 19th and the 20th values (that is, the two largest values) will be greater than or equal to at least $np = 20(0.95) = 19$ of the values and less than or equal to at least $n(1 - p) = 1$ value. The 95th percentile is thus the average of the 19th and 20th largest values.

Summing up, we have shown the following.

To find the sample $100p$ percentile of a data set of size n

1. Arrange the data in increasing order.
2. If np is not an integer, determine the smallest integer greater than np. The data value in that position is the sample $100p$ percentile.
3. If np is an integer, then the average of the values in positions np and $np + 1$ is the sample $100p$ percentile.

Example 3.9

Which data value is the sample 90th percentile when the sample size is (a) 8, (b) 16, and (c) 100?

Solution

(a) Since $0.9 \times 8 = 7.2$, which is not an integer, it follows that if the data are arranged from smallest to largest, then the sample 90th percentile value would be the 8th-smallest value (that is, the largest value).

(b) Since $0.9 \times 16 = 14.4$, which is not an integer, it follows that the sample 90th percentile would be the 15th-smallest value.

(c) Since $0.9 \times 100 = 90$ is an integer, the sample 90th percentile value is the average of the 90th and the 91st values when the data are arranged from smallest to largest. ■

Example 3.10

Table 3.1 lists the top 20 U.S. colleges and universities based on endowment assets. Using these data, find the

(a) Sample 90th percentile

(b) Sample 20th percentile

Table 3.1 Top 20 Colleges and Universities in Endowment Assets, 2005

	Institution	State	2005 Endowment Funds ($000)
1	Harvard University	MA	25,473,721
2	Yale University	CT	15,224,900
3	Stanford University	CA	12,205,000
4	University of Texas System	TX	11,610,997
5	Princeton University	NJ	11,206,500
6	Massachusetts Institute of Technology	MA	6,712,436
7	University of California	CA	5,221,916
8	Columbia University	NY	5,190,564
9	The Texas A&M University System and Foundations	TX	4,963,879
10	University of Michigan	MI	4,931,338
11	Emory University	GA	4,376,272
12	University of Pennsylvania	PA	4,369,782
13	Washington University	MO	4,268,415
14	Northwestern University	IL	4,215,275
15	University of Chicago	IL	4,137,494
16	Duke University	NC	3,826,153
17	Cornell University	NY	3,777,092
18	University of Notre Dame	IN	3,650,224
19	Rice University	TX	3,611,127
20	University of Virginia	VA	3,219,098

Solution

(a) Because the sample size is 20 and $20 \times 0.9 = 18$, the sample 90th percentile is the average of the 18th- and 19th-smallest values. Equivalently, it is the average of the 2nd- and 3rd-largest values. Hence,

$$\text{sample 90th percentile} = \frac{15{,}224{,}900 + 12{,}205{,}000}{2} = 13{,}714{,}950$$

That is, the sample 90th percentile of this data set is approximately $13.7 billion.

(b) Because $20 \times 0.2 = 4$, the sample 20th percentile is the average of the 4th- and 5th-smallest values, giving the result

$$\text{Sample 20th percentile} = \frac{3{,}777{,}092 + 3{,}826{,}153}{2} = 3{,}801{,}623$$ ■

The sample 25th percentile, 50th percentile, and 75th percentile are known as the *quartiles*.

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

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

Example 3.11

Find the sample quartiles for the following 18 data values, which represent the ordered values of a sample of scores from a league bowling tournament:

122, 126, 133, 140, 145, 145, 149, 150, 157, 162, 166, 175, 177, 177, 183, 188, 199, 212

Solution

Since $0.253 \times 18 = 4.5$, the sample 25th percentile is the fifth-smallest value, which is 145.

Since $0.50 \times 18 = 9$, the second quartile (or sample median) is the average of the 9th- and 10th-smallest values and so is

$$\frac{157 + 162}{2} = 159.5$$

Since $0.75 \times 18 = 13.5$, the third quartile is the 14th-smallest value, which is 177. ■

PROBLEMS

1. Seventy-five values are arranged in increasing order. How would you determine the sample
   (a) 80th percentile
   (b) 60th percentile
   (c) 30th percentile
   of this data set?
2. The following table gives the number of deaths of infants per 1,000 births in the 50 U.S. states in 2007. Use it to find the quartiles of the state infant death rates.

Number of Deaths of Infants per 1,000 Births and Total Infant Mortality

State Ranking			
1 Montana	4.5 (52 Total)	25 Arizona	6.7 (630 Total)
1 Vermont	4.5 (30 Total)	27 Kentucky	6.8 (378 Total)
3 Minnesota	4.7 (332 Total)	28 Florida	7.0 (1,537 Total)
4 Massachusetts	4.8 (380 Total)	29 Kansas	7.2 (284 Total)
5 Iowa	5.1 (195 Total)	29 Pennsylvania	7.2 (1,049 Total)
6 California	5.2 (2,811 Total)	31 Illinois	7.5 (1,349 Total)
6 Utah	5.2 (264 Total)	31 Missouri	7.5 (584 Total)
8 Rhode Island	5.3 (68 Total)	31 Virginia	7.5 (776 Total)
9 Connecticut	5.5 (233 Total)	34 Michigan	7.6 (984 Total)
9 Oregon	5.5 (251 Total)	34 West Virginia	7.6 (158 Total)
9 Washington	5.5 (451 Total)	36 Ohio	7.7 (1,143 Total)
12 New Hampshire	5.6 (81 Total)	37 Indiana	8.0 (700 Total)
12 New Jersey	5.6 (651 Total)	37 Oklahoma	8.0 (411 Total)
12 North Dakota	5.6 (46 Total)	39 South Dakota	8.2 (93 Total)
15 Hawaii	5.7 (104 Total)	40 Arkansas	8.3 (319 Total)
15 Maine	5.7 (79 Total)	41 Maryland	8.4 (630 Total)
17 Wisconsin	6.0 (420 Total)	42 Georgia	8.5 (1,181 Total)
18 New York	6.1 (1,518 Total)	43 Delaware	8.6 (98 Total)
19 Idaho	6.2 (139 Total)	43 Tennessee	8.6 (687 Total)
20 Colorado	6.3 (434 Total)	45 Alabama	8.7 (516 Total)
20 New Mexico	6.3 (179 Total)	46 North Carolina	8.8 (1,053 Total)
20 Texas	6.3 (2,407 Total)	46 Wyoming	8.8 (60 Total)
23 Nevada	6.4 (225 Total)	48 South Carolina	9.3 (525 Total)
24 Nebraska	6.6 (173 Total)	49 Mississippi	9.8 (420 Total)
25 Alaska	6.7 (69 Total)	50 Louisiana	10.5 (684 Total)

3. Consider a data set of n values $1, 2, 3, \ldots, n$. Find the value of the sample 95th percentile when
   (a) $n = 100$
   (b) $n = 101$

The following table gives the number of physicians and of dentists per 100,000 population for 12 midwestern states in 2000. Problems 4 and 5 are based on it.

4. For the physician's rates per 100,000 population, find the
   (a) Sample 40th percentile
   (b) Sample 60th percentile
   (c) Sample 80th percentile

State	Physician's rate	Dentist's rate
Ohio	188	56
Indiana	146	48
Illinois	206	61
Michigan	177	64
Wisconsin	177	70
Minnesota	207	70
Iowa	141	60
Missouri	186	55
North Dakota	157	55
South Dakota	129	54
Nebraska	162	71
Kansas	166	52

Source: American Medical Association, *Physician Characteristics and Distribution in the U.S.*

5. For the dentist's rates per 100,000 population, find the
   (a) Sample 90th percentile
   (b) Sample 50th percentile
   (c) Sample 10th percentile
6. Suppose the sample $100p$ percentile of a set of data is 120. If we add 30 to each data value, what is the new value of the sample $100p$ percentile?
7. Suppose the sample $100p$ percentile of a set of data is 230. If we multiply each data value by a positive constant c, what is the new value of the sample $100p$ percentile?
8. Find the sample 90th percentile of this data set:

$$75, 33, 55, 21, 46, 98, 103, 88, 35, 22, 29, 73, 37, 101,$$
$$121, 144, 133, 52, 54, 63, 21, 7$$

9. Use the table on page 96 to find the quartiles of 2006 traffic fatality rates (per 100 million vehicle miles) in the 50 states of the United States.
10. The following are the quartiles of a large data set:

$$\text{First quartile} = 35$$
$$\text{Second quartile} = 47$$
$$\text{Third quartile} = 66$$

Traffic Fatalities by State: 1990 to 2006
[For deaths within 30 days of the accident]

State	1990	2000	2005	2006	Fatality rate[1] 1990	Fatality rate[1] 2006	State	1990	2000	2005	2006	Fatality rate[1] 1990	Fatality rate[1] 2006
U.S...	**44,599**	**41,945**	**43,510**	**42,642**	**2.1**	**1.4**	MO....	1,097	1,157	1,257	1,096	2.2	1.6
AL	1,121	996	1,148	1,208	2.6	2.0	MT....	212	237	251	263	2.5	2.3
AK....	98	106	73	74	2.5	1.5	NE	262	276	276	269	1.9	1.4
AZ....	869	1,036	1,179	1,288	2.5	2.1	NV	343	323	427	432	3.4	2.0
AR	604	652	654	665	2.9	2.0	NH	158	126	166	127	1.6	0.9
CA....	5,192	3,753	4,333	4,236	2.0	1.3	NJ.....	886	731	747	772	1.5	1.0
CO....	544	681	606	535	2.0	1.1	NM....	499	432	488	484	3.1	1.9
CT....	385	341	278	301	1.5	1.0	NY....	2,217	1,460	1,434	1,456	2.1	1.0
DE	138	123	133	148	2.1	1.6	NC....	1,385	1,557	1,547	1,559	2.2	1.5
DC....	48	48	48	37	1.4	1.0	ND....	112	86	123	111	1.9	1.4
FL.....	2,891	2,999	3,518	3,374	2.6	1.7	OH....	1,638	1,366	1,321	1,238	1.8	1.1
GA....	1,562	1,541	1,729	1,693	2.2	1.5	OK....	641	650	803	765	1.9	1.6
HI.....	177	132	140	161	2.2	1.6	OR....	579	451	487	477	2.2	1.3
ID.....	244	276	275	267	2.5	1.8	PA....	1,646	1,520	1,616	1,525	1.9	1.4
IL.....	1,589	1,418	1,363	1,254	1.9	1.2	RI.....	84	80	87	81	1.1	1.0
IN.....	1,049	886	938	899	2.0	1.3	SC....	979	1,065	1,094	1,037	2.8	2.1
IA.....	465	445	450	439	2.0	1.4	SD....	153	173	186	191	2.2	2.1
KS....	444	461	428	468	1.9	1.6	TN....	1,177	1,307	1,270	1,287	2.5	1.8
KY....	849	820	985	913	2.5	1.9	TX....	3,250	3,779	3,536	3,475	2.1	1.5
LA....	959	938	963	982	2.5	2.2	UT	272	373	282	287	1.9	1.1
ME....	213	169	169	188	1.8	1.3	VT....	90	76	73	87	1.5	1.1
MD....	707	588	614	651	1.7	1.2	VA....	1,079	929	947	963	1.8	1.2
MA....	605	433	441	430	1.3	0.8	WA....	825	631	649	630	1.8	1.1
MI.....	1,571	1,382	1,129	1,085	1.9	1.0	WV....	481	411	374	410	3.1	2.0
MN....	566	625	559	494	1.5	0.9	WI.....	769	799	815	724	1.7	1.2
MS....	750	949	931	911	3.1	2.2	WY....	125	152	170	195	2.1	2.1

[1]Deaths per 100 million vehicle miles traveled.
Source: U.S. National Highway Traffic Safety Administration, *Traffic Safety Facts*, annual. See <http://www-nrd.nhtsa.dot.gov/CATS/Index.aspx>.

(a) Give an interval in which approximately 50 percent of the data lie.
(b) Give a value which is greater than approximately 50 percent of the data.
(c) Give a value such that approximately 25 percent of the data values are greater than it.

11. A symmetric data set has its median equal to 40 and its third quartile equal to 55. What is the value of the first quartile?

3.4 SAMPLE MODE

Another indicator of central tendency is the *sample mode*, which is the data value that occurs most frequently in the data set.

■ Example 3.12

The following are the sizes of the last 8 dresses sold at a women's boutique:

$$8, 10, 6, 4, 10, 12, 14, 10$$

What is the sample mode?

Solution

The sample mode is 10, since the value of 10 occurs most frequently. ■

If no single value occurs most frequently, then all the values that occur at the highest frequency are called *modal values*. In such a situation we say that there is no unique value of the sample mode.

■ Example 3.13

The ages of 6 children at a day care center are

$$2, 5, 3, 5, 2, 4$$

What are the modal values of this data set?

Solution

Since the ages 2 and 5 both occur most frequently, both 2 and 5 are modal values. ■

■ Example 3.14

The following frequency table gives the values obtained in 30 throws of a die.

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

It is easy to pick out the modal value from a frequency table, since it is just that value having the largest frequency.

For these data, find the

(a) Sample mode
(b) Sample median
(c) Sample mean

Solution

(a) Since the value 4 appears with the highest frequency, the sample mode is 4.
(b) Since there are 30 data values, the sample median is the average of the 15th- and 16th-smallest values. Since the 15th-smallest value is 3 and the 16th-smallest is 4, the sample median is 3.5.
(c) The sample mean is

$$\bar{x} = \frac{1 \cdot 6 + 2 \cdot 4 + 3 \cdot 5 + 4 \cdot 8 + 5 \cdot 3 + 6 \cdot 4}{30} = \frac{100}{30} \approx 3.333$$ ■

PROBLEMS

1. Match each statement in the left-hand column with the correct data set from the right-hand column.

1. Sample mode is 9	A: 5, 7, 8, 10, 13, 14
2. Sample mean is 9	B: 1, 2, 5, 9, 9, 15
3. Sample median is 9	C: 1, 2, 9, 12, 12, 18

2. Using the data from Example 2.2, find the sample mode of the winning Masters Golf Tournament scores.
3. Using data concerning the first 100 students in App. A, find the sample mode for
(a) Weight
(b) Blood pressure
(c) Cholesterol

4. Suppose you want to guess the salary of a bank vice president whom you have just met. If you want to have the greatest chance of being correct to the nearest $1000, would you rather know the sample mean, the sample median, or the sample mode of the salaries of bank vice presidents?
5. Construct a data set for which the sample mean is 10, the sample median is 8, and the sample mode is 6.
6. If the sample mode of the data $x_i, i = 1, \ldots, n$, is equal to 10, what is the sample mode of the data $y_i = 2x_i + 5, i = 1, \ldots, n$?
7. Joggers use a quarter-mile track around an athletic field. In a sample of 17 joggers, 1 did 2 loops, 4 did 4 loops, 5 did 6 loops, 6 did 8 loops, and 1 did 12 loops.
   (a) What is the sample mode of the number of loops run by these joggers?
   (b) What is the sample mode of the distances run by these joggers?
8. The sample mean, sample median, and sample mode of the first 99 values of a data set of 198 values are all equal to 120. If the sample mean, median, and mode of the final 99 values are all equal to 100, what can you say about the sample mean of the entire data set? What can you say about the sample median? What about the sample mode?

3.5 SAMPLE VARIANCE AND SAMPLE STANDARD DEVIATION

Whereas so far we have talked about statistics that measure the central tendency of a data set, we have not yet considered ones that measure its spread or variability. For instance, although the following data sets A and B have the same sample mean and sample median, there is clearly more spread in the values of B than in those of A.

$$\boldsymbol{A}\text{: } 1, 2, 5, 6, 6 \quad \boldsymbol{B}\text{: } -40, 0, 5, 20, 35$$

One way of measuring the variability of a data set is to consider the deviations of the data values from a central value. The most commonly used central value for this purpose is the sample mean. If the data values are $x_1, \ldots, x_n$ and the sample mean is $\bar{x} = \sum_{i=1}^{n} x_i/n$ then the deviation of the value x_i from the sample mean is $x_i - \bar{x}, i = 1, \ldots, n$.

One might suppose that a natural measure of the variability of a set of data would be the average of the deviations from the mean. However, as we have shown in Sec. 3.2, $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$. That is, the sum of the deviations from the sample mean is always equal to 0, and thus the average of the deviations from the sample mean must also be 0. However, after some additional reflection it should be clear that we really do not want to allow the positive and the negative deviations to

cancel. Instead, we should be concerned about the individual deviations without regard to their signs. This can be accomplished either by considering the absolute values of the deviations or, as turns out to be more useful, by considering their squares.

The sample variance is a measure of the "average" of the squared deviations from the sample mean. However, for technical reasons (which will become clear in Chap. 8) this "average" divides the sum of the n squared deviations by the quantity $n - 1$, rather than by the usual value n.

Definition *The* sample variance, *call it* s^2, *of the data set* $x_i, \ldots, x_n$ *having sample mean* $\bar{x} = \left(\sum_{i=1}^{n} x_i\right) n$ *is defined by*

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

Example 3.15

Find the sample variance of data set A.

Solution

It is determined as follows:

x_i	1	2	5	6	6
$\bar{x}$	4	4	4	4	4
$x_i - \bar{x}$	−3	−2	1	2	2
$(x_i - \bar{x})^2$	9	4	1	4	4

Hence, for data set A,

$$s^2 = \frac{9 + 4 + 1 + 4 + 4}{4} = 5.5$$ ■

Example 3.16

Find the sample variance for data set B.

Solution

The sample mean for data set B is also $\bar{x} = 4$. Therefore, for this set, we have

x_i	−40	0	5	20	35
$x_i - \bar{x}$	−44	−4	1	16	31
$(x_i - \bar{x})^2$	1936	16	1	256	961

Thus,

$$s^2 = \frac{3170}{4} = 792.5$$ ■

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

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 \tag{3.2}$$

■ Example 3.17

Check that identity (3.2) holds for data set *A*.

Solution

Since $n = 5$ and $\bar{x} = 4$,

$$\sum_{i=1}^{5} x_i^2 - n\bar{x}^2 = 1 + 4 + 25 + 36 + 36 - 5(16) = 102 - 80 = 22$$

From Example 3.15,

$$\sum_{i=1}^{5}(x_i - \bar{x})^2 = 9 + 4 + 1 + 4 + 4 = 22$$

and so the identity checks out. ■

Suppose that we add a constant c to each of the data values $x_1, \ldots, x_n$ to obtain the new data set $y_1, \ldots, y_n$, where

$$y_i = x_i + c$$

To see how this affects the value of the sample variance, recall from Sec. 3.2 that

$$\bar{y} = \bar{x} + c$$

and so

$$y_i - \bar{y} = x_i + c - (\bar{x} + c) = x_i - \bar{x}$$

That is, the y deviations are equal to the x deviations, and therefore their sums of squares are equal. Thus, we have shown the following useful result.

The sample variance remains unchanged when a constant is added to each data value.

The preceding result can often be used in conjunction with the algebraic identity (3.2) to greatly reduce the time it takes to compute the sample variance.

Example 3.18

The following data give the yearly numbers of law enforcement officers killed in the United States over 10 years:

$$164, 165, 157, 164, 152, 147, 148, 131, 147, 155$$

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

Solution

Rather than working directly with the given data, let us subtract the value 150 from each data item. (That is, we are adding $c = -150$ to each data value.) This results in the new data set

$$14, 15, 7, 14, 2, -3, -2, -19, -3, 5$$

Its sample mean is

$$\bar{y} = \frac{14 + 15 + 7 + 14 + 2 - 3 - 2 - 19 - 3 + 5}{10} = 3.0$$

The sum of the squares of the new data is

$$\sum_{i=1}^{10} y_i^2 = 14^2 + 15^2 + 7^2 + 14^2 + 2^2 + 3^2 + 2^2 + 19^2 + 3^2 + 5^2 = 1078$$

Therefore, using the algebraic identity (3.2) shows that

$$\sum_{i=1}^{10} (y_i - \bar{y})^2 = 1078 - 10(9) = 988$$

Hence, the sample variance of the revised data, which is equal to the sample variance of the original data, is

$$s^2 = \frac{988}{9} \approx 109.78$$ ■

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

Definition *The quantity* s, *defined by*

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

is called the sample standard deviation.

The sample standard deviation is measured in the same units as the original data. That is, for instance, if the data are in feet, then the sample variance will be expressed in units of square feet and the sample standard deviation in units of feet.

If each data value $x_i, i = 1, \ldots, n$, is multiplied by a constant c to obtain the new data set

$$y_i = cx_i \qquad i = 1, \ldots, n$$

then the sample variance of the y data is the sample variance of the x data multiplied by c^2. That is,

$$s_y^2 = c^2 s_x^2$$

where s_y^2 and s_x^2 are the sample variances of the new and old data sets, respectively. Taking the square root of both sides of the preceding equation shows that the standard deviation of the y data is equal to the absolute value of c times the standard deviation of the x data, or

$$s_y = |c| s_x$$

Another indicator of the variability of a data set is the interquartile range, which is equal to the third minus the first quartile. That is, roughly speaking, the interquartile range is the length of the interval in which the middle half of the data values lie.

Example 3.19

The Miller Analogies Test is a standardized test that is taken by a variety of students applying to graduate and professional schools. Table 3.2 presents some of the percentile scores on this examination for students, classified according to the graduate fields they are entering. For instance, Table 3.2 states that the median grade of students in the physical sciences is 68, whereas it is 49 for those applying to law school.

Determine the interquartile ranges of the scores of students in the five specified categories.

Table 3.2 Selected Percentiles on the Miller Analogies Test for Five Categories of Students

Percentile	Physical sciences	Medical school	Social sciences	Languages and literature	Law school
99	93	92	90	87	84
90	88	78	82	80	73
75	80	71	74	73	60
50	68	57	61	59	49
25	55	45	49	43	37

Solution

Since the interquartile range is the difference between the 75th and the 25th sample percentiles, it follows that its value is

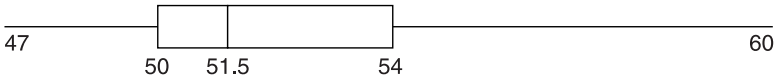
$80 - 55 = 25$ for scores of physical science students
$71 - 45 = 26$ for scores of medical school students
$74 - 49 = 25$ for scores of social science students
$73 - 43 = 30$ for scores of language and literature students
$60 - 37 = 23$ for scores of law school students ■

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 following frequency table gives the starting salaries of a sample of 42 graduating seniors of a liberal arts college.

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

The salaries go from a low of 47 to a high of 60. The value of the first quartile (equal to the value of the 11th smallest on the list) is 50; the value of the second

quartile (equal to the average of the 21st- and 22nd-smallest values) is 51.5; and the value of the third quartile (equal to the value of the 32nd smallest on the list) is 54. The box plot for this data set is as follows.



A box plot

PROBLEMS

1. The following data give the per capita consumption of milk in the years from 1983 to 1987. The data are from the U.S. Department of Agriculture, *Food Consumption, Prices, and Expenditures*, annual.

Year	Amount (in gallons per capita)
1983	26.3
1984	26.2
1985	26.4
1986	26.3
1987	25.9

   Find the sample mean and the sample variance of this set.

2. You are given these data sets:

   A: 66, 68, 71, 72, 72, 75 *B*: 2, 5, 9, 10, 10, 16

   (a) Which one appears to have the larger sample variance?
   (b) Determine the sample variance of data set *A*.
   (c) Determine the sample variance of data set *B*.

3. The Masters Golf Tournament and the U.S. Open are the two most prestigious golf tournaments in the United States. The Masters is always played on the Augusta National golf course, whereas the U.S. Open is played on different courses in different years. As a result, one might expect the sample variance of the winning scores in the U.S. Open to be higher than that of the winning scores in the Masters. To check whether this is so, we have collected the winning scores in both tournaments for 1981 to 1990.

	Winning score									
Tournament	**1981**	**1982**	**1983**	**1984**	**1985**	**1986**	**1987**	**1988**	**1989**	**1990**
U.S. Open	273	282	280	276	279	279	277	278	278	280
Masters	280	284	280	277	282	279	285	281	283	278

(a) Compute the sample variance of the winning scores in the U.S. Open tournament.
(b) Compute the sample variance of the winning scores in the Masters tournament.

The following table gives the numbers of physicians and dentists in Japan in the even-numbered years between 1984 and 2000. Problems 4 and 5 are based on this table.

Number of Physicians and Dentists (1984–2000)

	Physicians	**Dentists**
1984	173,452	61,283
1986	183,129	64,904
1988	193,682	68,692
1990	203,797	72,087
1992	211,498	75,628
1994	220,853	79,091
1996	230,297	83,403
1998	236,933	85,669
2000	243,201	88,410

4. Just by eyeballing, estimate the ratio of the sample variance of the yearly number of physicians to the sample variance of the yearly number of dentists.
5. Find the actual value of the ratio in problem 4.
6. An individual needing automobile insurance requested quotes from 10 different insurers for identical coverage and received the following values (amounts are annual premiums in dollars):

720, 880, 630, 590, 1140, 908, 677, 720, 1260, 800

Find
(a) The sample mean
(b) The sample median
(c) The sample standard deviation

The following table gives the 2008 populations for each U.S. state and territory. Problems 7, 8, and 9 refer to this data.

State	Population (2008)	State	Population (2008)
California	36,756,666	Oklahoma	3,642,361
Texas	24,326,974	Connecticut	3,501,252
New York	19,490,297	Iowa	3,002,555
Florida	18,328,340	Mississippi	2,938,618
Illinois	12,901,563	Arkansas	2,855,390
Pennsylvania	12,448,279	Kansas	2,802,134
Ohio	11,485,910	Utah	2,736,424
Michigan	10,003,422	Nevada	2,600,167
Georgia	9,685,744	New Mexico	1,984,356
North Carolina	9,222,414	West Virginia	1,814,468
New Jersey	8,682,661	Nebraska	1,783,432
Virginia	7,769,089	Idaho	1,523,816
Washington	6,549,224	Maine	1,316,456
Arizona	6,500,180	New Hampshire	1,315,809
Massachusetts	6,497,967	Hawaii	1,288,198
Indiana	6,376,792	Rhode Island	1,050,788
Tennessee	6,214,888	Montana	967,440
Missouri	5,911,605	Delaware	873,092
Maryland	5,633,597	South Dakota	804,194
Wisconsin	5,627,967	Alaska	686,293
Minnesota	5,220,393	North Dakota	641,481
Colorado	4,939,456	Vermont	621,270
Alabama	4,661,900	*District of Columbia*	591,833
South Carolina	4,479,800	Wyoming	532,668
Louisiana	4,410,796	*Guam*	173,456
Kentucky	4,269,245	*US Virgin Islands*	108,448
Puerto Rico	3,954,037	*Northern Mariana Islands*	84,546
Oregon	3,790,060	*American Samoa*	57,291

7. Find the sample variance of the populations of the first 17 locales.
8. Find the sample variance of the populations of the next 17 locales.
9. Find the sample variance of the populations of the final 17 locales.
10. If s^2 is the sample variance of the data $x_i, i = 1, \ldots, n$, what is the sample variance of the data $ax_i + b, i = 1, \ldots, n$, when a and b are given constants?
11. Compute the sample variance and sample standard deviation of the following data sets:
    (a) 1, 2, 3, 4, 5
    (b) 6, 7, 8, 9, 10
    (c) 11, 12, 13, 14, 15
    (d) 2, 4, 6, 8, 10
    (e) 10, 20, 30, 40, 50

12. On the U.S. side of the U.S.–Canada border, temperatures are measured in degrees Fahrenheit, whereas on the Canadian side they are measured in degrees Celsius (also called Centigrade). Suppose that during the month of January the sample mean of the temperatures, as recorded on the U.S. side of the border, was 40°F with a sample variance of 12.

Use the formula for converting a Fahrenheit temperature to a Celsius temperature

$$C = \frac{5}{9}(F - 32)$$

to find

(a) The sample mean recorded by the Canadians
(b) The sample variance recorded by the Canadians

13. Compute the sample mean and sample variance of the systolic blood pressures of the first 50 students of the data set of App. A. Now do the same with the last 50 students of this data set. Compare your answers. Comment on the results of this comparison. Do you find it surprising?

14. If s is the sample standard deviation of the data $x_i, i = 1, \ldots, n$, what is the sample standard deviation of $ax_i + b, i = 1, \ldots, n$? In this problem, a and b are given constants.

15. The following table gives the number of motorcycle retail sales in Japan for 8 different years. Use it to find the sample standard deviation of the number of motorcycle sales in the 8 years.

Year	2001	2002	2003	2004	2005	2006	2007	2008
Motorcycle sales (in thousands)	751	771	760	700	707	700	685	522

Source: Motorcycle Industry Council.

16. Find the sample standard deviation of the data set given by the following frequency table:

Value	Frequency	Value	Frequency
3	1	5	3
4	2	6	2

17. The following data represent the acidity of 40 successive rainfalls in the state of Minnesota. The acidity is measured on a pH scale, which varies from 1 (very acidic) to 7 (neutral).

3.71, 4.23, 4.16, 2.98, 3.23, 4.67, 3.99, 5.04, 4.55, 3.24, 2.80, 3.44, 3.27, 2.66, 2.95, 4.70, 5.12, 3.77, 3.12, 2.38, 4.57, 3.88, 2.97, 3.70, 2.53, 2.67, 4.12, 4.80, 3.55, 3.86, 2.51, 3.33, 3.85, 2.35, 3.12, 4.39, 5.09, 3.38, 2.73, 3.07

(a) Find the sample standard deviation.
(b) Find the range.
(c) Find the interquartile range.

18. Consider the following two data sets:

A: 4.5, 0, 5.1, 5.0, 10, 5.2 B: 0.4, 0.1, 9, 0, 10, 9.5

(a) Determine the range for each data set.
(b) Determine the sample standard deviation for each data set.
(c) Determine the interquartile range for each data set.

3.6 NORMAL DATA SETS AND THE EMPIRICAL RULE

Many of the large data sets one encounters in practice have histograms that are similar in shape. These histograms are often symmetric about their point of highest frequency and then decrease on both sides of this point in a bell-shaped fashion. Such data sets are said to be *normal*, and their histograms are called *normal histograms*.

Definition *A data set is said to be* normal *if a histogram describing it has the following properties:*

1. *It is highest at the middle interval.*
2. *Moving from the middle interval in either direction, the height decreases in such a way that the entire histogram is bell-shaped.*
3. *The histogram is symmetric about its middle interval.*

Figure 3.2 shows the histogram of a normal data set.

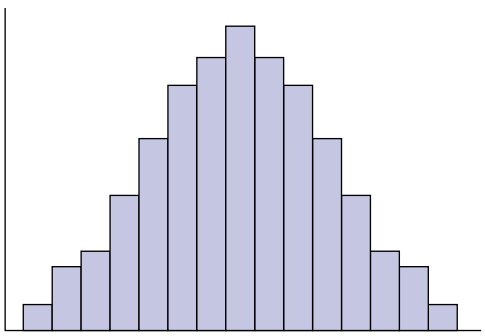


FIGURE 3.2
Histogram of a normal data set.

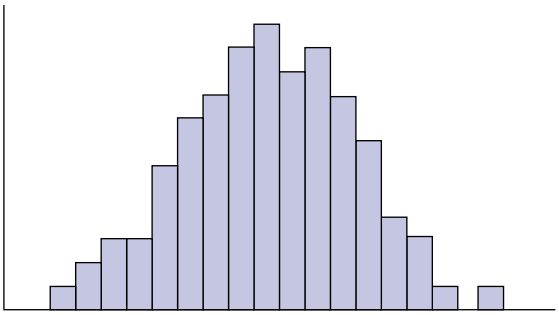


FIGURE 3.3
Histogram of an approximately normal data set.

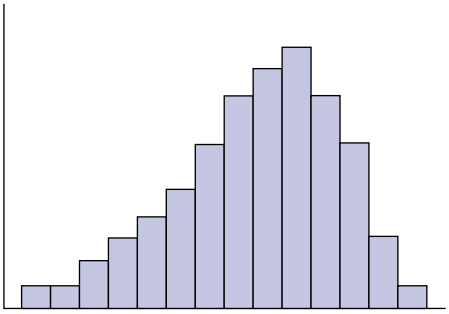


FIGURE 3.4
Histogram of a data set skewed to the left.

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, the histogram given in Fig. 3.3 is from an approximately normal data set, whereas the ones presented in Figs. 3.4 and 3.5 are not (since each is too nonsymmetric). Any data set that is not approximately symmetric about its sample median is said to be *skewed*. It is called *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 Fig. 3.4 is skewed to the left, and the one of Fig. 3.5 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.

Suppose that $\bar{x}$ and s are the sample mean and sample standard deviation, respectively, 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 $\bar{x}$.

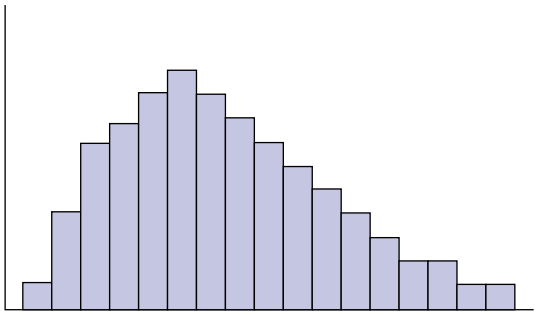


FIGURE 3.5
Histogram of a data set skewed to the right.

Empirical Rule

If a data set is approximately normal with sample mean $\bar{x}$ and sample standard deviation s, then the following 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 3.20

The scores of 25 students on a history examination are listed on the following stem-and-leaf plot.

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

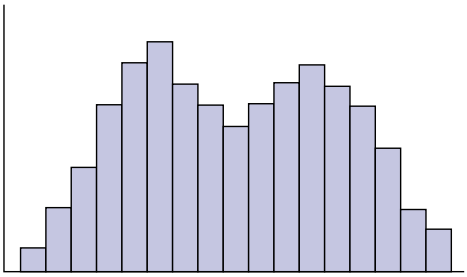


FIGURE 3.6
Histogram of a bimodal data set.

By standing this figure on its side (or, equivalently, by turning the textbook), we can see that the corresponding histogram is approximately normal. Use it to assess the empirical rule.

Solution

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

$$\bar{x} = 73.68 \quad \text{and} \quad s = 12.80$$

The empirical rule states that approximately 68 percent of the data values are between $\bar{x} - s = 60.88$ and $\bar{x} + s = 86.48$. Since 17 of the observations actually fall within 60.88 and 86.48, the actual percentage is $100(17/25) = 68$ percent. Similarly, the empirical rule states that approximately 95 percent of the data are between $\bar{x} - 2s = 48.08$ and $\bar{x} + 2s = 96.28$, whereas, in actuality, 100 percent of the data fall in this range. ■

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 Fig. 3.6 is bimodal.

Since a stem-and-leaf plot can be regarded as a histogram lying on its side, it is useful in showing us whether a data set is approximately normal.

■ Example 3.21

The following is the stem-and-leaf plot of the weights of 200 members of a health club.

24	9
23	
22	1
21	7
20	2, 2, 5, 5, 6, 9, 9, 9
19	0, 0, 0, 0, 0, 1, 1, 2, 4, 4, 5, 8
18	0, 1, 1, 2, 2, 2, 3, 4, 4, 4, 5, 5, 5, 6, 6, 6, 6, 7, 9, 9, 9
17	1, 1, 1, 2, 3, 3, 4, 4, 4, 5, 5, 6, 6, 6, 6, 7, 7, 7, 7, 9
16	0, 0, 1, 1, 1, 1, 2, 4, 5, 5, 6, 6, 8, 8, 8, 8
15	0, 1, 1, 1, 1, 1, 1, 5, 5, 5, 5, 6, 6, 6, 7, 7, 8, 9
14	0, 0, 0, 1, 2, 3, 4, 5, 6, 7, 7, 7, 8, 9, 9
13	0, 0, 0, 1, 1, 1, 2, 2, 2, 2, 2, 3, 3, 4, 5, 5, 6, 6, 6, 6, 7, 7, 8, 8, 8, 9, 9, 9
12	1, 1, 1, 2, 2, 2, 3, 4, 4, 5, 5, 6, 6, 6, 6, 6, 6, 6, 6, 7, 7, 7, 7, 8, 8, 9, 9, 9
11	0, 1, 1, 2, 2, 2, 2, 3, 3, 4, 4, 5, 5, 6, 9, 9
10	0, 2, 3, 3, 3, 4, 4, 5, 7, 7, 8
9	0, 0, 9
8	6

By standing it on its side, we see that its histogram does not appear to be approximately normal. However, it is important to note that these data consist of the weights of all members of the health club, both female and male. Since these are clearly separate populations with regard to weight, it makes sense to consider the data for each gender separately. We will now do so.

It turns out that these 200 data values are the weights of 97 women and 103 men. Separating the data for women and men results in the stem-and-leaf plots in Figs. 3.7 and 3.8.

As we can see from the figures, the separated data for each sex appear to be approximately normal. Let us calculate $\bar{x}_w$, s_w, $\bar{x}_m$, and s_m, the sample mean and sample standard deviation for, respectively, the women and the men.

This calculation yields

$$\bar{x}_w = 125.70 \quad \bar{x}_m = 174.69$$
$$s_w = 15.58 \quad s_m = 21.23$$

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

FIGURE 3.7

Weights of 97 female health club members.

24	9
23	
22	1
21	7
20	2, 2, 5, 5, 6, 9, 9, 9
19	0, 0, 0, 0, 0, 1, 1, 2, 4, 4, 5, 8
18	0, 1, 1, 2, 2, 2, 3, 4, 4, 4, 5, 5, 5, 6, 6, 6, 6, 7, 9, 9, 9
17	1, 1, 1, 2, 3, 3, 4, 4, 4, 5, 5, 6, 6, 6, 6, 7, 7, 7, 7, 9
16	0, 1, 1, 1, 1, 2, 4, 5, 6, 6, 8, 8, 8, 8
15	1, 1, 1, 5, 5, 5, 6, 6, 6, 7, 7, 8, 9
14	0, 5, 7, 7, 8, 9
13	0, 1, 2, 3, 7
12	9

FIGURE 3.8
Weights of 103 male health club members.

A further corroboration of the approximate normality of the two sets of separated data is provided by noting the similar values in each set of the sample mean and sample median. The sample median of the women's weights is the 49th-smallest data value, which equals 126, whereas for the men's data the sample median is the 52nd-smallest data value, which equals 174. These are quite close to the two sample means, whose values are 125.7 and 174.69.

Given the values of the sample mean and sample standard deviation, it follows from the empirical rule that approximately 68 percent of the women will weigh between 110.1 and 141.3 and approximately 95 percent of the men will weigh between 132.2 and 217.2. The actual percentages from Figs. 3.7 and 3.8 are

$$100 \times \frac{68}{97} = 70.1 \quad \text{and} \quad 100 \times \frac{101}{103} = 98.1$$

■

PROBLEMS

1. The daily numbers of animals treated at a certain veterinarian clinic over a 24-day period are as follows:

22, 17, 19, 31, 28, 29, 21, 33, 36, 24, 15, 28, 25, 28, 22,
27, 33, 19, 25, 28, 26, 20, 30, 32

(a) Plot these data in a histogram.
(b) Find the sample mean.

(c) Find the sample median.
(d) Is this data set approximately normal?

Historical Perspective

Quetelet and How the Normal Curve Uncovered Fraud


(North Wind Picture Archives)

Adolphe Quetelet

The Belgian social scientist and statistician Adolphe Quetelet was a great believer in the hypothesis that most data sets relating to human measurements are normal. In one study he measured the chests of 5738 Scottish soldiers, plotted the resulting data in a histogram, and concluded that it was normal.

In a later study Quetelet used the shape of the normal histogram to uncover evidence of fraud in regard to draft conscripts to the French army. He studied data concerning the heights of a huge sample of 100,000 conscripts. Plotting the data in a histogram—with class intervals of 1 inch—he found that, with the exception of three class intervals around 62 inches, the data appeared to be normal. In particular, there were fewer values in the interval from 62 to 63 inches and slightly more in the intervals from 60 to 61 and from 61 to 62 inches than would have occurred with a perfect normal fit of the data. Trying to figure out why the normal curve did not fit as well as he had supposed it would, Quetelet discovered that 62 inches was the minimum height required for soldiers in the French army. Based on this and his confidence in the widespread applicability of normal data, Quetelet concluded that some conscripts whose heights were slightly above 62 inches were "bending their knees" to appear shorter so as to avoid the draft.

For 50 years following Quetelet, that is, roughly from 1840 to 1890, it was widely believed that most data sets from homogeneous populations (that is, data that were not obviously a mixture of different populations) would appear to be normal if the sample size were sufficiently large. Whereas present-day statisticians have become somewhat skeptical about this claim, it is quite common for a data set to appear to come from a normal population. This phenomenon, which often appears in data sets originating in either the biological or the physical sciences, is partially explained by a mathematical result known as the *central limit theorem*. Indeed, the central limit theorem (studied in Chap. 7) will in itself explain why many data sets originating in the physical sciences are approximately normal. To explain why biometric data (that is, data generated by studies in biology) often appear to be normal, we will use what was originally an empirical observation noted by Francis Galton but that nowadays has a sound scientific explanation, called *regression to the mean*. Regression to the mean, in conjunction with the central limit theorem and the passing of many generations, will yield our explanation as to why a biometric data set is often normal. The explanation will be presented in Chap. 12.

2. The following data give the injury rates per 100,000 worker-hours for a sample of 20 semiconductor firms:

1.4, 2.4, 3.7, 3.1, 2.0, 1.9, 2.5, 2.8, 2.2, 1.7, 3.1, 4.0, 2.2, 1.8,
2.6, 3.6, 2.9, 3.3, 2.0, 2.4

(a) Plot the data in a histogram.
(b) Is the data set roughly symmetric?
(c) If the answer to (b) is no, is it skewed to the left or to the right?
(d) If the answer to (b) is yes, is it approximately normal?

The following table gives the 2006 per capita consumption of milk in various countries. Problems 3 and 4 refer to this table.

Per Capita Consumption of Milk and Milk Products in Various Countries, 2006 data

Country	Liquid milk drinks (litres)
Finland	183.9
Sweden	145.5
Ireland	129.8
Netherlands	122.9
Norway	116.7
Spain (2005)	119.1
Switzerland	112.5
United Kingdom (2005)	111.2
Australia (2005)	106.3
Canada (2005)	94.7
European Union (25 countries)	92.6
Germany	92.3
France	92.2
New Zealand (2005)	90.0
United States	83.9
Austria	80.2
Greece	69.0
Argentina (2005)	65.8
Italy	57.3
Mexico	40.7
China (2005)	8.8

Source: International Dairy Federation, Bulletin 423/2007.

3. Find the sample mean and sample median of the milk consumption data set.
4. Plot the milk consumption data in a stem and leaf plot. Is the data set approximately normal?

5. The following represent the times (in minutes) it took 22 newly hired workers to complete a standardized task:

166, 82, 175, 181, 169, 177, 180, 185, 159, 164, 170, 149, 188, 173, 170, 164, 158, 177, 173, 175, 190, 172

   (a) Find the sample mean.
   (b) Find the sample median.
   (c) Plot the data in a histogram.
   (d) Is this data set approximately normal?

6. The following data give the age at inauguration of all 43 presidents of the United States.

President	Age at inauguration	President	Age at inauguration
1. Washington	57	23. B. Harrison	55
2. J. Adams	61	24. Cleveland	55
3. Jefferson	57	25. McKinley	54
4. Madison	57	26. T. Roosevelt	42
5. Monroe	58	27. Taft	51
6. J.Q. Adams	57	28. Wilson	56
7. Jackson	61	29. Harding	55
8. Van Buren	54	30. Coolidge	51
9. W. Harrison	68	31. Hoover	54
10. Tyler	51	32. F. Roosevelt	51
11. Polk	49	33. Truman	60
12. Taylor	64	34. Eisenhower	62
13. Fillmore	50	35. Kennedy	43
14. Pierce	48	36. L. Johnson	55
15. Buchanan	65	37. Nixon	56
16. Lincoln	52	38. Ford	61
17. A. Johnson	56	39. Carter	52
18. Grant	46	40. Reagan	69
19. Hayes	54	41. G. H. W. Bush	64
20. Garfield	49	42. Clinton	46
21. Arthur	50	43. G. W. Bush	54
22. Cleveland	47	44. Obama	47

   (a) Find the sample mean and sample standard deviation of this data set.
   (b) Draw a histogram for the given data.
   (c) Do the data appear to be approximately normal?

(d) If the answer to (c) is yes, give an interval that you would expect to contain approximately 95 percent of the data observations.
(e) What percentage of the data lies in the interval given in part (d)?

7. For the data on the weights of female health club members presented in Fig. 3.7, the sample mean and sample standard deviation were computed to be 125.70 and 15.58, respectively. Based on the shape of Fig. 3.7 and these values, approximate the proportion of the women whose weight is between 94.54 and 156.86 pounds. What is the actual proportion?
8. A sample of 36 male coronary patients yielded the following data concerning the ages at which they suffered their first heart attacks.

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

(a) Determine $\bar{x}$ and s.
(b) From the shape of the stem-and-leaf plot, what percentage of data values would you expect to be between $\bar{x} - s$ and $\bar{x} + s$? Between $\bar{x} - 2s$ and $\bar{x} + 2s$?
(c) Find the actual percentages for the intervals given in (b).

9. If the histogram is skewed to the right, which statistic will be larger—the sample mean or the sample median? (*Hint*: If you are not certain, construct a data set that is skewed to the right and then calculate the sample mean and sample median.)
10. The following data are the ages of a sample of 36 victims of violent crime in a large eastern city:

25, 16, 14, 22, 17, 20, 15, 18, 33, 52, 70, 38, 18, 13, 22, 27, 19, 23,
33, 15, 13, 62, 21, 57, 66, 16, 24, 22, 31, 17, 20, 14, 26, 30, 18, 25

(a) Determine the sample mean.
(b) Find the sample median.
(c) Determine the sample standard deviation.
(d) Does this data set appear to be approximately normal?
(e) What proportion of the data lies within 1 sample standard deviation of the sample mean?
(f) Compare your answer in (e) to the approximation provided by the empirical rule.

The following table lists the 2002 per capita income for the 50 states. Problems 11 to 13 refer to it.

11. Using the data on the first 25 states,

- **(a)** Plot the data in a histogram.
- **(b)** Compute the sample mean.
- **(c)** Compute the sample median.
- **(d)** Compute the sample variance.
- **(e)** Are the data approximately normal?
- **(f)** Use the empirical rule to give an interval which should contain approximately 68 percent of the observations.
- **(g)** Use the empirical rule to give an interval which should contain approximately 95 percent of the observations.
- **(h)** Determine the actual proportion of observations in the interval specified in (f).
- **(i)** Determine the actual proportion of observations in the interval specified in (g).

12. Repeat Prob. 11, this time using the data on the final 25 states.

13. Repeat Prob. 11, this time using all the data in the table.

Personal Income per Capita in Constant (1996) Dollars, 2002

State	Income	Rank	State	Income	Rank
United States	27,857	(X)	Kansas	26,237	26
Alabama	22,624	43	Kentucky	23,030	39
Alaska	28,947	14	Louisiana	22,910	41
Arizona	23,573	38	Maine	24,979	33
Arkansas	21,169	49	Maryland	32,680	4
California	29,707	10	Massachusetts	35,333	3
Colorado	29,959	9	Michigan	27,276	18
Connecticut	38,450	1	Minnesota	30,675	7
Delaware	29,512	12	Mississippi	20,142	50
Florida	26,646	23	Missouri	26,052	27
Georgia	25,949	28	Montana	22,526	45
Hawaii	27,011	20	Nebraska	26,804	22
Idaho	22,560	44	Nevada	27,172	19
Illinois	30,075	8	New Hampshire	30,912	6
Indiana	25,425	32	New Jersey	35,521	2
Iowa	25,461	31	New Mexico	21,555	47
New York	32,451	5	Tennessee	24,913	35
North Carolina	24,949	34	Texas	25,705	30
North Dakota	24,293	36	Utah	21,883	46
Ohio	26,474	25	Vermont	26,620	24
Oklahoma	23,026	40	Virginia	29,641	11
Oregon	25,867	29	Washington	29,420	13

(*Continued*)

(*Continued*)

State	Income	Rank	State	Income	Rank
Pennsylvania	28,565	15	West Virginia	21,327	48
Rhode Island	28,198	16	Wisconsin	26,941	21
South Carolina	22,868	42	Wyoming	27,530	17
South Dakota	24,214	37			

Note: When states share the same rank, the next lower rank is omitted. Because of rounded data, states may have identical values shown, but different ranks.

3.7 SAMPLE CORRELATION COEFFICIENT

Consider the data set of paired values $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. In this section we will present a statistic, called the *sample correlation coefficient*, that measures the degree to which larger x values go with larger y values and smaller x values go with smaller y values.

The data in Table 3.3 represent the average daily number of cigarettes smoked (the x variable) and the number of free radicals (the y variable), in a suitable unit, found in the lungs of 10 smokers. (A free radical is a single atom of oxygen. It is believed to be potentially harmful because it is highly reactive and has a strong tendency to combine with other atoms within the body.) Figure 3.9 shows the scatter diagram for these data.

From an examination of Fig. 3.9 we see that when the number of cigarettes is high, there tends to be a large number of free radicals, and when the number of

Table 3.3 Cigarette Smoking and Free Radicals

Person	Number of cigarettes smoked	Free radicals
1	18	202
2	32	644
3	25	411
4	60	755
5	12	144
6	25	302
7	50	512
8	15	223
9	22	183
10	30	375

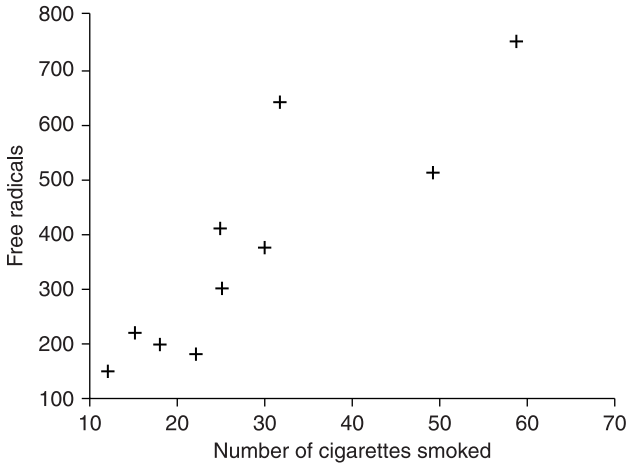


FIGURE 3.9

Cigarettes smoked versus number of free radicals.

Table 3.4 Pulse Rate and Years of School Completed

	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

cigarettes smoked is low, there tends to be a small number of free radicals. In this case, we say that there is a *positive correlation* between these two variables.

We are also interested in determining the strength of the relationship between a pair of variables in which large values of one variable tend to be associated with small values of the other. For instance, the data of Table 3.4 represent the years of schooling (variable x) and the resting pulse rate in beats per minute (variable y) of 10 individuals. A scatter diagram of this data is presented in Fig. 3.10. From Fig. 3.10 we see that higher numbers of years of schooling tend to be associated with lower resting pulse rates and that lower numbers of years of schooling tend to be associated with the higher resting pulse rates. This is an example of a *negative correlation*.

To obtain a statistic that can be used to measure the association between the individual values of a paired set, suppose the data set consists of the paired values $(x_i, y_i), i = 1, \ldots, n$. Let $\bar{x}$ and $\bar{y}$ denote the sample mean of the x values and the sample mean of the y values, respectively. For data pair i, consider $x_i - \bar{x}$ the

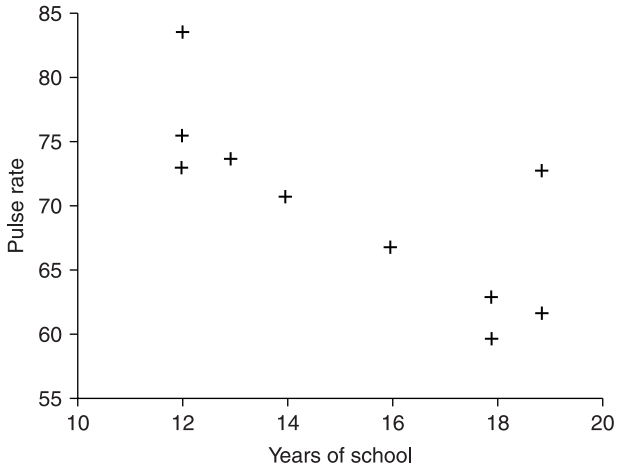


FIGURE 3.10
Scatter diagram of years in school and pulse rate.

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 and 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. Since 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.

The same logic also implies that when large values of one of the variables tend to go along with small values of the other, 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.

We now list some of the properties of the sample correlation coefficient.

1. The sample correlation coefficient r is always between -1 and $+1$.
2. The sample correlation coefficient r will equal $+1$ if, for some constant a,

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

   where b is a positive constant.
3. The sample correlation coefficient r will equal -1 if, for some constant a,

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

   where b is a negative constant.
4. If r is the sample correlation coefficient for the data x_i, y_i, $i = 1, \ldots, n$, then for any constants a, b, c, d, it is also the sample correlation coefficient for the data

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

   provided that b and d have the same sign (that is, provided that $bd \geq 0$).

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 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 degrees Fahrenheit or Celsius.

For computational purposes, the following is a convenient formula for the sample correlation coefficient.

Computational Formula for r

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

Example 3.22

The following table gives the U.S. per capita consumption of whole milk (x) and of low-fat milk (y) in three different years.

	Per capita consumption (gallons)		
	1980	1984	1984
Whole milk (x)	17.1	14.7	12.8
Low-fat milk (y)	10.6	11.5	13.2

Source: U.S. Department of Agriculture, *Food Consumption, Prices, and Expenditures.*

Find the sample correlation coefficient r for the given data.

Solution

To make the computation easier, let us first subtract 12.8 from each of the x values and 10.6 from each of the y values. This gives the new set of data pairs:

	i		
	1	**2**	**3**
x_i	4.3	1.9	0
y_i	0	0.9	2.6

Now,

$$\bar{x} = \frac{4.3 + 1.9 + 0}{3} = 2.0667$$

$$\bar{y} = \frac{0 + 0.9 + 2.6}{3} = 1.1667$$

$$\sum_{i=1}^{3} x_i y_i = (1.9)(0.9) = 1.71$$

$$\sum_{i=1}^{3} x_i^2 = (4.3)^2 + (1.9)^2 = 22.10$$

$$\sum_{i=1}^{3} y_i^2 = (0.9)^2 + (2.6)^2 = 7.57$$

Thus,

$$r = \frac{1.71 - 3(2.0667)(1.1667)}{\sqrt{[22.10 - 3(2.0667)^2][7.57 - 3(1.1667)^2]}} = -0.97$$

Therefore, our three data pairs exhibit a very strong negative correlation between consumption of whole and of low-fat milk.

For small data sets such as in Example 3.22, the sample correlation coefficient can be easily obtained by hand. However, for large data sets this computation can become tedious, and a calculator or statistical software is useful. ■

■ Example 3.23

Compute the sample correlation coefficient of the data of Table 3.3, which relates the number of cigarettes smoked to the number of free radicals found in a person's lungs.

Solution

The number of pairs is 10. The pairs are as follows:

18, 202

32, 644

25, 411

60, 755

12, 144

25, 302

50, 512

15, 223

22, 183

30, 375

A calculation shows that the sample correlation coefficient is 0.8759639. ■

The large value of the sample correlation coefficient indicates a strong positive correlation between the number of cigarettes a person smokes and the number of free radicals in that person's lungs.

Example 3.24

Compute the sample correlation coefficient of the data of Table 3.4, which relates a person's resting pulse rate to the number of years of school completed.

Solution

The pairs are as follows:

12, 73

16, 67

13, 74

18, 63

19, 73

12, 84

18, 60

19, 62

12, 76

14, 71

The sample correlation coefficient is -0.763803.

The large negative value of the sample correlation coefficient indicates that, for the data set considered, a high pulse rate tends to be associated with a small number of years spent in school and a low pulse rate tends to be associated with a large number of years spent in school. ■

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 about 0.8 means that the linear relation is relatively strong; although there is no straight line that passes through all the data points, there is one that is "close" to them all. A value of $|r|$ around 0.3 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

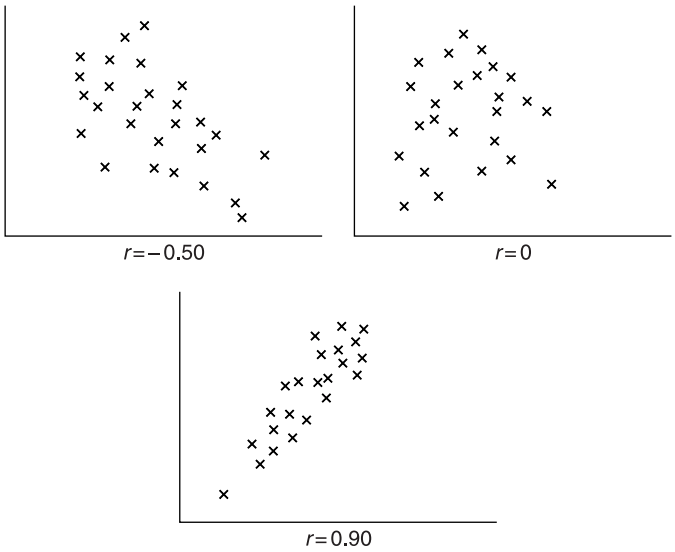


FIGURE 3.11
Sample correlation coefficients.

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 3.11 displays scatter diagrams for data sets with various values of r.

Historical Perspective


(Bettmann)
Francis Galton

The development of the concept and utility of the sample correlation coefficient involved the efforts of four of the great men of statistics. The original concept was due to Francis Galton, who was trying to study the laws of inheritance from a quantitative point of view. As such, he wanted to be able to quantify the degree to which characteristics of an offspring relate to those of its parents. This led him to name and define a form of the sample correlation coefficient that differs somewhat from the one presently in use. Although originally it was meant to be used to assess the hereditary influence of a parent on an offspring, Galton later realized that the sample correlation coefficient presented a method of assessing the interrelation between any two variables.

Although Francis Galton was the founder of the field of biometrics—the quantitative study of biology—its acknowledged leader, at least after 1900, was Karl Pearson. After the Royal Society of London passed a resolution in 1900 stating that it would no longer accept papers that applied mathematics to the study

of biology, Pearson, with financial assistance from Galton, founded the statistical journal *Biometrika*, which still flourishes today. The form of the sample correlation coefficient that is presently in use (and that we have presented) is due to Karl Pearson and was originally called *Pearson's product-moment correlation coefficient*.

The probabilities associated with the possible values of the sample correlation coefficient r were discovered, in the case where the data pairs come from a normal population, by William Gosset. There were, however, some technical errors in his derivations, and these were subsequently corrected in a paper by Ronald Fisher.

PROBLEMS

1. Explain why the sample correlation coefficient for the data pairs

   $$(121, 360), (242, 362), (363, 364)$$

   is the same as that for the pairs

   $$(1, 0), (2, 2), (3, 4)$$

   which is the same as that for the pairs

   $$(1, 0), (2, 1), (3, 2)$$

2. Compute the sample correlation coefficient for the data pairs in Prob. 1.

Statistics In Perspective

Correlation Measures Association, Not Causation

The results of Example 3.24 indicated 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 perhaps it is not knowledge that is making the difference but rather that people who have had more education tend to end up in jobs that allow them more time for exercise and good nutrition. Probably the strong negative correlation between years in school and resting pulse rate results from a combination of these as well as other underlying factors.

3. The following data represent the IQ scores of 10 mothers and their eldest daughters.

Mother's IQ	Daughter's IQ
135	121
127	131
124	112
120	115
115	99
112	118
104	106
96	89
94	92
85	90

(a) Draw a scatter diagram.
(b) Guess at the value of the sample correlation coefficient r.
(c) Compute r.
(d) What conclusions can you draw about the relationship between the mother's and daughter's IQs?

4. The following is a sampling of 10 recently released first-time federal prisoners. The data give their crime, their sentence, and the actual time that they served.

Number	Crime	Sentence (months)	Time served (months)
1	Drug abuse	44	24
2	Forgery	30	12
3	Drug abuse	52	26
4	Kidnapping	240	96
5	Income tax fraud	18	12
6	Drug abuse	60	28
7	Robbery	120	52
8	Embezzlement	24	14
9	Robbery	60	35
10	Robbery	96	49

Draw a scatter diagram of the sentence time versus time actually served. Compute the sample correlation coefficient. What does this say about the relationship between the length of a sentence and the time actually served?

5. Using the data of Prob. 4, determine the sample correlation coefficient of the sentence time and the proportion of that time actually served.

What does this say about the relationship between the length of a sentence and the proportion of this time that is actually served?

6. The following data refer to the number of adults in prison and on parole in 12 midwestern states. The data are in thousands of adults.

State	In prison	On parole
Illinois	18.63	11.42
Indiana	9.90	2.80
Iowa	2.83	1.97
Kansas	4.73	2.28
Michigan	17.80	6.64
Minnesota	2.34	1.36
Missouri	9.92	4.53
Nebraska	1.81	0.36
North Dakota	0.42	0.17
Ohio	20.86	6.51
South Dakota	1.05	0.42
Wisconsin	5.44	3.85

(a) Draw a scatter diagram.
(b) Determine the sample correlation coefficient between the number of adults in state prison and on parole in that state.
(c) Fill in the missing word. States having a large prison population tend to have a(n)_____number of individuals on parole.

7. The following data relate the number of criminal cases filed in various U.S. cities to the percentage of those cases that result in a plea of guilty.

City	Percentage of cases resulting in a guilty plea	Number of cases filed
San Diego, CA	73	11,534
Dallas, TX	72	14,784
Portland, OR	62	3,892
Chicago, IL	41	35,528
Denver, CO	68	3,772
Philadelphia, PA	26	13,796
Lansing, MI	68	1,358
St. Louis, MO	63	3,649
Davenport, IA	60	1,312
Tallahassee, FL	50	2,879
Salt Lake City, UT	61	2,745

Determine the sample correlation coefficient between the number of cases filed and the percentage of guilty pleas. What can you say about the degree of association between these two variables for these data?

8. The following table gives yearly per capita soft drink consumption (in litres) and the yearly per capita milk consumption (in kg) for a variety of countries. Use it to find the sample correlation coefficient between soft drink and milk consumption.

Per capita soft drink and milk consumption

Country	soft drink	milk
United States	216	254
Australia	100	233
Switzerland	81	308
France	37	256
United Kingdom	97	230
The Netherlands	96	329
New Zealand	84	210
Germany	72	314
Italy	50	239
Japan	22	68

9. The following table lists per capita income data both for the U.S. and for residents of the state of Colorado for each of the years from 1992 to 2007. Use it to compute the sample correlation coefficient between U.S. and Colorado per capita income.

Annual Per Capita Personal Income		
	United States	Colorado
1992	$20,854	$21,109
1993	$21,346	$22,054
1994	$22,172	$23,004
1995	$23,076	$24,226
1996	$24,175	$25,570
1997	$25,334	$26,846
1998	$26,883	$28,784
1999	$27,939	$30,492
2000	$29,845	$33,361
2001	$30,574	$34,438
2002	$30,821	$33,956
2003	$31,504	$33,989
2004	$33,123	$35,523
2005	$34,757	$37,600
2006	$36,714	$39,491
2007	$38,611	$41,042

10. The following data give the numbers of physicians and dentists, per 100,000 population, in the United States for six different years.

	1980	1981	1982	1983	1985	1986	2001
Physicians	211	217	222	228	237	246	253
Dentists	54	54	55	56	57	57	59

Source: *Health Resources Statistics, annual.*

(a) Show that the number of physicians and the number of dentists are positively correlated for these years.

(b) Do you think that a large value of one of these variables by itself causes a large value of the other? If not, how would you explain the reason for the positive correlation?

The following table gives the death rates by selected causes in different countries. It will be used in Probs. 11 to 13.

Death Rates per 100,000 Population by Selected Causes and Countries

				Malignant neoplasm of—				
Country	**Year**	**Ischemic heart disease**	**Cerebro-vascular disease**	**Lung, trachea, bronchus**	**Stomach**	**Female breast**	**Bronchitis, emphysema, asthma**	**Chronic lever disease and cirrhosis**
United States	1984	218.1	60.1	52.7	6.0	31.9	8.3	12.9
Australia	1985	230.9	95.6	41.0	10.1	30.0	16.9	8.7
Austria	1986	155.1	133.2	34.3	20.7	31.6	22.3	26.6
Belgium	1984	120.6	95.0	55.9	14.7	36.8	22.6	12.4
Bulgaria	1985	245.9	254.5	30.6	24.2	21.5	28.6	16.2
Canada	1985	200.6	57.5	50.6	9.0	34.5	9.7	10.1
Czechoslovakia	1985	289.4	194.3	51.3	22.4	27.3	33.8	19.6
Denmark	1985	243.8	73.4	52.2	10.9	39.7	37.1	12.2
Finland	1986	259.8	105.0	36.4	17.3	23.9	19.8	8.8
France	1985	76.0	79.7	32.2	10.8	27.1	11.7	22.9
Hungary	1986	240.1	186.5	55.0	25.9	31.2	43.8	42.1
Italy	1983	128.9	121.9	42.1	23.9	28.9	30.9	31.5
Japan	1986	41.9	112.8	24.9	40.7	8.1	12.2	14.4
Netherlands	1985	164.6	71.1	56.3	15.6	38.2	17.8	5.5
New Zealand	1985	250.5	98.4	42.0	11.2	37.7	25.8	4.8
Norway	1985	208.5	88.6	26.3	14.4	25.9	18.2	6.9
Poland	1986	109.4	75.3	47.2	24.2	21.1	33.4	12.0

(*Continued*)

Country	Year	Ischemic heart disease	Cerebro-vascular disease	Malignant neoplasm of— Lung, trachea, bronchus	Stomach	Female breast	Bronchitis, emphysema, asthma	Chronic lever disease and cirrhosis
Portugal	1986	76.6	216.4	18.7	26.5	22.6	17.8	30.0
Spain	1981	79.0	133.9	26.0	19.7	19.0	19.1	23.3
Sweden	1985	244.7	73.0	23.2	12.5	26.0	14.3	6.4
Switzerland	1986	112.0	65.6	36.6	12.0	36.6	17.5	10.4
United Kingdom:								
England and Wales	1985	247.6	104.5	57.2	15.2	41.9	24.2	4.8
Scotland	1986	288.0	128.4	68.7	14.9	41.2	14.8	7.3
West Germany	1986	159.5	100.4	34.6	18.3	32.6	26.1	19.3

Source: World Health Organization, *World Health Statistics*.

In doing Probs. 11 to 13, use all the data if you are running either Program 3-2 or a statistical package. If you are working with a hand calculator, use only the data relating to the first seven countries.

11. Find the sample correlation coefficient between the death rates of ischemic heart disease and of chronic liver disease.

12. Find the sample correlation coefficient between the death rates of stomach cancer and of female breast cancer.

13. Find the sample correlation coefficient between the death rates of lung cancer and of bronchitis, emphysema, and asthma.

14. In a well-publicized experiment, a University of Pittsburgh researcher enlisted the cooperation of public school teachers in Boston in obtaining a baby tooth from each of their pupils. These teeth were then sawed open and analyzed for lead content. The lead content of each tooth was plotted against the pupil's IQ test score. A strong negative correlation resulted between the amount of lead in the teeth and the IQ scores. Newspapers headlined this result as "proof" that lead ingestion results in decreased scholastic aptitude.

(a) Does this conclusion necessarily follow?

(b) Offer some other possible explanations.

15. A recent study has found a strong positive correlation between the cholesterol levels of young adults and the amounts of time they spend watching television.

(a) Would you have expected such a result? Why?

(b) Do you think that watching television causes higher cholesterol levels?

(c) Do you think that having a high cholesterol level makes a young adult more likely to watch television?
(d) How would you explain the results of the study?

16. An analysis relating the number of points scored and fouls committed by basketball players in the Pacific Ten conference has established a strong positive correlation between these two variables. The analyst has gone on record as claiming that this verifies the hypothesis that offensive-minded basketball players tend to be very aggressive and so tend to commit a large number of fouls. Can you think of a simpler explanation for the positive correlation? (*Hint*: Think in terms of the average number of minutes per game that a player is on the court.)
17. A *New England Journal of Medicine* study published in October 1993 found that people who have guns in their homes for protection are 3 times more likely to be murdered than those with no guns in the home. Does this prove that an individual's chance of being murdered is increased when he or she purchases a gun to keep at home? Explain your answer.
18. 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?
19. 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.
20. A recent study (reported in the March 10, 2009 NY Times) yielded a negative correlation between the age of the father and the results of cognitive tests given to the infant at ages 8 months, 4 years old, and 7 years old. Although the differences in scores between those infants having younger and older fathers was slight the authors of the study called the findings "unexpectedly startling." (On the other hand there was a positive correlation between a mother's age and the cognitive test scores.) Discuss the potential difficulties in interpreting the results of this study.

KEY TERMS

Statistic: A numerical quantity whose value is determined by the data.

Sample mean: The arithmetic average of the values in a data set.

Deviation: The difference between the individual data values and the sample mean. If x_i is the ith data value and $\bar{x}$ is the sample mean, then $x_i - \bar{x}$ is called the *ith deviation*.

Sample median: The middle value of an ordered set of data. For a data set of n values, the sample median is the $(n+1)/2$-smallest value when n is odd and the average of the $n/2$- and $n/2+1$-smallest values when n is even.

Sample 100p percentile: 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 of the data are greater than or equal to it. If two data values satisfy this criterion, then it is the average of them.

First quartile: The sample 25th percentile.

Second quartile: The sample 50th percentile, which is also the sample median.

Third quartile: The sample 75th percentile.

Sample mode: The data value that occurs most frequently in a data set.

Sample variance: The statistic s^2, defined by

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

It measures the average of the squared deviations.

Sample standard deviation: The positive square root of the sample variance.

Range: The largest minus the smallest data value.

Interquartile range: The third quartile minus the first quartile.

Normal data set: One whose histogram is symmetric about its middle interval and decreases on both sides of the middle in a bell-shaped manner.

Skewed data set: One whose histogram is not symmetric about its middle interval. It is said to be 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.

Bimodal data set: One whose histogram has two local peaks or humps.

Sample correlation coefficient: For the set of paired values $x_i, y_i, i = 1, \ldots, n$, it is defined by

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

where $\bar{x}$ and s_x are, respectively, the sample mean and the sample standard deviation of the x values, and similarly for $\bar{y}$ and s_y. A value of r near $+1$ indicates that larger x values tend to be paired with larger y values and smaller x values tend to be paired with smaller y values. A value near -1 indicates that larger x values

tend to be paired with smaller y values and smaller x values tend to be paired with larger y values.

SUMMARY

We have seen three different statistics which describe the center of a data set: the sample mean, sample median, and sample mode.

The sample mean of the data $x_1 \ldots, x_n$ is defined by

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

and is a measure of the center of the data.

If the data are specified by the frequency table

Value	Frequency
x_1	f_1
x_2	f_2
$\vdots$	$\vdots$
x_k	f_k

then the sample mean of the $n = \sum_{i=1}^{k} f_i$ data values can be expressed as

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

A useful identity is

$$\sum_{i=1}^{n} (x_i - \bar{x}) = 0$$

The sample median is the middle value when the data are arranged from smallest to largest. If there are an even number of data points, then it is the average of the two middle values. It is also a measure of the center of the data set.

The sample mode is that value in the data set that occurs most frequently.

Suppose a data set of size n is arranged from smallest to largest. If np is not an integer, then the sample $100p$ percentile is the value whose position is the smallest integer larger than np. If np is an integer, then the sample $100p$ percentile is the average of the values in positions np and $np + 1$.

The sample 25th percentile is the *first quartile*. The sample 50th percentile (which is equal to the sample median) is called the *second quartile*, and the sample 75th percentile is called the *third quartile*.

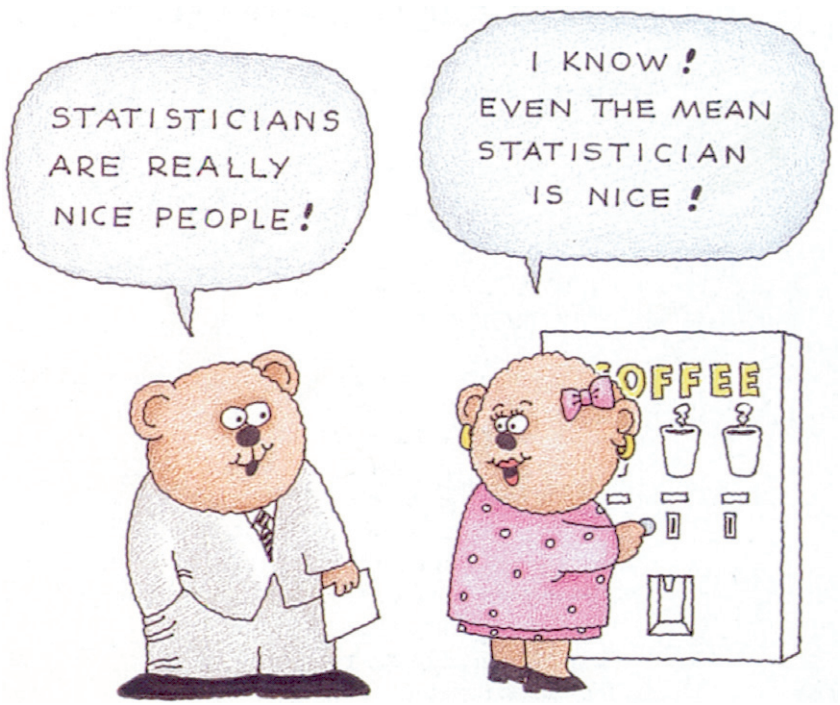
The sample variance s^2 is a measure of the spread in the data and is defined by

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

where n is the size of the set. Its square root s is called the *sample standard deviation*, and it is measured in the same units as the data.

The following identity is useful for computing the sample variance by using pencil and paper or a hand calculator.

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



Program 3-1 will compute the sample mean, sample variance, and sample standard deviation of any set of data.

Another statistic that describes the spread of the data is the *range*, the difference between the largest and smallest data values.

Normal data sets will have their sample mean and sample median approximately equal. Their histograms are symmetric about the middle interval and exhibit a bell shape.

The sample correlation coefficient r measures the degree of association between two variables. Its value is between -1 and $+1$. A value of r near $+1$ indicates that when one of the variables is large, the other one also tends to be large and that when one of them is small, the other also tends to be small. A value of r near -1 indicates that when one of the variables is large, the other one tends to be small.

A large value of $|r|$ indicates a strong association between the two variables. Association does not imply causation.

REVIEW PROBLEMS

1. Construct a data set that is symmetric about 0 and contains
   (a) Four distinct values
   (b) Five distinct values
   (c) In both cases, compute the sample mean and sample median.
2. The following stem-and-leaf plot records the diastolic blood pressure of a sample of 30 men.

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

   (a) Compute the sample mean $\bar{x}$.
   (b) Compute the sample median.
   (c) Compute the sample mode.
   (d) Compute the sample standard deviation s.
   (e) Do the data appear to be approximately normal?
   (f) What proportion of the data values lies between $\bar{x} + 2s$ and $\bar{x} - 2s$?
   (g) Compare the answer in part (f) to the one prescribed by the empirical rule.
3. The following data are the median ages of residents in each of the 50 states of the United States:

29.3	27.7	30.4	31.1	28.5
32.1	28.0	31.3	26.6	25.8
25.9	33.0	31.5	30.0	28.4
24.9	31.6	26.6	25.4	29.2
29.3	27.9	31.8	31.5	30.3

28.5	29.3	26.6	31.2	32.1
31.4	30.1	27.0	28.5	27.6
28.9	29.4	30.5	31.2	29.4
29.3	30.1	28.8	27.9	30.4
32.3	30.4	25.8	27.1	26.9

(a) Find the median of these ages.
(b) Is this necessarily the median age of all people in the United States? Explain.
(c) Find the quartiles.
(d) Find the sample 90th percentile.

4. Use Table 3.2 in Example 3.19 to fill in the answers.
(a) To have one's score be among the top 10 percent of all physical science students, it must be at least___.
(b) To have one's score be among the top 25 percent of all social science students, it must be at least___.
(c) To have one's score be among the bottom 50 percent of all medical students, it must be less than or equal to___.
(d) To have one's score be among the middle 50 percent of all law school students, it must be between___ and___.

5. The number of violent offenses per 100,000 population is given here for each of the 50 states. Is this data set approximately normal?

Violent Crime per 100,000 Population, 2002

State	Rate	Rank	State	Rate	Rank
United States	495	(X)	Illinois	621	8
Alabama	444	21	Iowa	286	36
Alaska	563	12	Kansas	377	24
Arizona	553	13	Kentucky	279	38
Arkansas	424	22	Louisiana	662	6
California	593	10	Maine	108	48
Colorado	352	27	Maryland	770	2
Connecticut	311	33	Massachusetts	484	18
Delaware	599	9	Michigan	540	14
Florida	770	2	Minnesota	268	40
Georgia	459	20	Mississippi	343	31
Hawaii	262	41	Missouri	539	15
Indiana	357	26	Montana	352	27
Idaho	255	42	Nebraska	314	32

(*Continued*)

(*Continued*)

State	Rate	Rank	State	Rate	Rank
Nevada	638	7	South Carolina	822	1
New Hampshire	161	47	South Dakota	177	46
New Jersey	375	25	Tennessee	717	5
New Mexico	740	4	Texas	579	11
New York	496	17	Utah	237	43
North Carolina	470	19	Vermont	107	49
North Dakota	78	50	Virginia	291	35
Ohio	351	29	Washington	345	30
Oklahoma	503	16	West Virginia	234	44
Oregon	292	34	Wisconsin	225	45
Pennsylvania	402	23	Wyoming	274	39
Rhode Island	285	37			

Note: Violent crime refers to violent offenses known to the police, which includes murder, forcible rape, robbery, and aggravated assault. When states share the same rank, the next lower rank is omitted. Because of rounded data, states may have identical values shown but different ranks.

6. The following data represent the birth weights at an inner-city hospital in a large eastern city:

2.4, 3.3, 4.1, 5.0, 5.1, 5.2, 5.6, 5.8, 5.9, 5.9, 6.0, 6.1, 6.2, 6.3,
6.3, 6.4, 6.4, 6.5, 6.7, 6.8, 7.2, 7.4, 7.5, 7.5, 7.6, 7.6, 7.7, 7.8,
7.8, 7.9, 7.9, 8.3, 8.5, 8.8, 9.2, 9.7, 9.8, 9.9, 10.0, 10.3, 10.5

(a) Plot this in a stem-and-leaf diagram.
(b) Find the sample mean $\bar{x}$.
(c) Find the sample median.
(d) Find the sample standard deviation s.
(e) What proportion of the data lies within $\bar{x} \pm 2s$?
(f) Do the data appear to be approximately normal?
(g) If your answer to (f) is yes, what would you have estimated, based on your answers to (b) and (d), for (e)?

***7.** Let a and b be constants. Show that if $y_i = a + bx_i$ for $i = 1, \ldots, n$, then r, the sample correlation coefficient of the data pairs $x_i, y_i, i = 1, \ldots, n$, is given by
(a) $r = 1$ when $b > 0$
(b) $r = -1$ when $b < 0$
(*Hint*: Use the definition of r, not its computational formula.)

8. The following data are taken from the book *Researches on the Probability of Criminal and Civil Verdicts*, published in 1837 by the French mathematician and probabilist Simeon Poisson. The book emphasized

legal applications of probability. The data refer to the number of people accused and convicted of crimes in France from 1825 to 1830.

Year	Number accused	Number convicted
1825	6652	4037
1826	6988	4348
1827	6929	4236
1828	7396	4551
1829	7373	4475
1830	6962	4130

(a) Determine the sample mean and sample median of the number accused.
(b) Determine the sample mean and sample median of the number convicted.
(c) Determine the sample standard deviation of the number accused.
(d) Determine the sample standard deviation of the number convicted.
(e) Would you expect the number accused and the number convicted to have a positive or a negative sample correlation coefficient?
(f) Determine the sample correlation coefficient of the number of accused and number of convicted.
(g) Determine the sample correlation coefficient between the number accused and the percentage of these who are convicted.
(h) Draw scatter diagrams for parts (f) and (g).
(i) Guess at the value of the sample correlation coefficient between the number of convicted and the percentage convicted.
(j) Draw a scatter diagram for the variables in (i).
(k) Determine the sample correlation coefficient for the variables in part (i).

9. Recent studies have been inconclusive about the connection between coffee consumption and coronary heart disease. If a study indicated that consumers of large amounts of coffee appeared to have a greater chance of suffering heart attacks than did drinkers of moderate amounts or drinkers of no coffee at all, would this necessarily "prove" that excessive coffee drinking leads to an increased risk of heart attack? What other explanations are possible?

10. Recent studies have indicated that death rates for married middle-aged people appear to be lower than for single middle-aged people. Does this mean that marriage tends to increase one's life span? What other explanations are possible?

11. A June 9, 1994, article in *The New York Times* noted a study showing that years with low inflation rates tend to be years with high average-productivity increases. The article claimed that this supported the Federal Reserve Board's claim that a low rate of inflation tends to result in an increase in productivity. Do you think the study provides strong evidence for this claim? Explain your answer.

12. The following table gives the 2008 medicare enrollment as a percentage of the total population for each of the 50 states and the District of Columbia.

(a) Find the sample mean of these data values.

(b) Is your answer to part (a) necessarily equal to the percentage of the entire population that is enrolled in medicare? Why or why not?

Alabama	17%	Missouri	16%
Alaska	8%	Montana	16%
Arizona	13%	Nebraska	15%
Arkansas	18%	Nevada	13%
California	12%	New Hampshire	15%
Colorado	12%	New Jersey	15%
Connecticut	15%	New Mexico	15%
Delaware	16%	New York	15%
District of Columbia	13%	North Carolina	15%
Florida	17%	North Dakota	16%
Georgia	12%	Ohio	16%
Hawaii	15%	Oklahoma	16%
Idaho	14%	Oregon	15%
Illinois	15%	Pennsylvania	18%
Indiana	15%	Rhode Island	17%
Iowa	17%	South Carolina	16%
Kansas	15%	South Dakota	16%
Kentucky	17%	Tennessee	16%
Louisiana	15%	Texas	11%
Maine	19%	Utah	10%
Maryland	13%	Vermont	17%
Massachusetts	16%	Virginia	14%
Michigan	15%	Washington	14%
Minnesota	14%	West Virginia	20%
Mississippi	16%	Wisconsin	15%
		Wyoming	14%

13. A sample of size $n + m$ consists of numerical values from n men and m women. If $\bar{x}_w$ is the sample mean of the women's values, and $\bar{x}_m$ is the sample mean of the men's values what is the sample mean of the entire sample?

14. 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?
15. The following are the number of traffic deaths in a sample of states, both for 2007 and 2008. Plot a scatter diagram and find the sample correlation coefficient for the data pairs.

2007 and 2008 Traffic Fatalities per State

State	2007	2008
WY	149	159
IL	1248	1044
MA	434	318
NJ	724	594
MD	615	560
OR	452	414
WA	568	504
FL	3221	2986
UT	291	271
NH	129	139