CHAPTER 13

Now in its 122nd edition and weighing over 2 pounds, the *Statistical Abstract of the United States* is a compilation of facts and figures taken from the U.S. census. In it you'll find many items, including what Americans owe and what kinds of pets Americans have (in 2001 36.1% of households had dogs, 31.6% had cats, and 4.6% had birds).

STATISTICS

Benjamin Disraeli (1804–1881), once Prime Minister of Britain, said that there are three kinds of lies: "Lies, damned lies, and statistics." Do numbers lie? Numbers are, after all, the foundation of all statistical information. The "lie" comes in when, either intentionally or carelessly, a number is used in such a way as to lead us to a conclusion that is unjustified or incorrect.

The first large-scale survey was commissioned in 1086 by William the Conqueror of England to provide a basis for taxation. The first modern census, for use as a basis for government representation, was taken in the United States in August 1790. A census has been taken in the United States every 10 years since then, and it is now used for many purposes.

Today there are methods by which statistics obtained from a small sample are used to represent a much larger population. In fact, a sample as small as 1600 may be used to predict the outcome of a national election. Information gathering is conducted by a variety of people, including medical researchers, scientists, advertisers, and political pollsters. You are likely to find examples of statistics every day on the front page of your newspaper.

When evaluating statistical information, remember that you need to judge the sampling methods as well as the numbers given. Ask yourself who conducted the study and whether they have a bias, how large the sample was and whether it was representative, and where the study appeared and whether there was an opposing side. Numbers may not lie, but they can be manipulated and misinterpreted.

13.1 SAMPLING TECHNIQUES

DID YOU KNOW

Tune in Tomorrow

The A. C. Nielsen Company, which has been measuring the viewing population of TV shows for more than 50 years, uses a sample of 5100 households in the United States to draw conclusions about more than 93 million viewers. An electronic measurement system, called the People Meter, is placed on each TV in the sample household. Nielsen uses the People Meter to measure what program is being tuned in and who is watching. Each household member is assigned a personal viewing button on the People Meter to keep track of which channels he or she watches and for how long. Nielsen then computes the rating of the show, using the data obtained from their sample. The People Meter is used to collect audience information for broadcast and cable networks, nationally syndicated programs, Spanish-language networks, and satellite distributors. Due in part to the claim by ABC, CBS, and NBC that its statistics are unreliable, Nielsen may soon begin using a new video recorder to monitor viewing habits.

Statistics is the art and science of gathering, analyzing, and making inferences (predictions) from numerical information obtained in an experiment. This numerical information is referred to as ***data***. The use of statistics, originally associated with numbers gathered for governments, has grown significantly and is now applied in all walks of life.

Governments use statistics to estimate the amount of unemployment and the cost of living. Thus, statistics has become an indispensable tool in attempting to regulate the economy. In psychology and education, the statistical theory of tests and measurements has been developed to compare achievements of individuals from diverse places and backgrounds. Another use of statistics with which we are all familiar is the public opinion poll. Newspapers and magazines carry the results of different polls on topics ranging from the president's popularity to the number of cans of soda consumed. In recent years, these polls have attained a high degree of accuracy. The A. C. Nielsen rating is a public opinion poll that determines the country's most and least watched TV shows. Statistics is used in scores of other professions; in fact, it is difficult to find one that does not depend on some aspect of statistics.

Statistics is divided into two main branches: descriptive and inferential. ***Descriptive statistics*** is concerned with the collection, organization, and analysis of data. ***Inferential statistics*** is concerned with making generalizations or predictions from the data collected.

Probability and statistics are closely related. Someone in the field of probability is interested in computing the chance of occurrence of a particular event when all the possible outcomes are known. A statistician's interest lies in drawing conclusions about possible outcomes through observations of only a few particular events.

If a probability expert and a statistician find identical boxes, the probability expert might open the box, observe the contents, replace the cover, and proceed to compute the probability of randomly selecting a specific object from the box. The statistician might select a few items from the box without looking at the contents and make a prediction as to the total contents of the box.

The entire contents of the box constitute the ***population***. A population consists of all items or people of interest. The statistician often uses a subset of the population, called a ***sample***, to make predictions concerning the population. It is important to understand the difference between a population and a sample. A population includes *all* items of interest. A sample includes *some* of the items in the population.

When a statistician draws a conclusion from a sample, there is always the possibility that the conclusion is incorrect. For example, suppose that a jar contains 90 blue marbles and 10 red marbles, as shown in Fig. 13.1. If the statistician selects a random sample of five marbles from the jar and all are blue, he or she may wrongly conclude

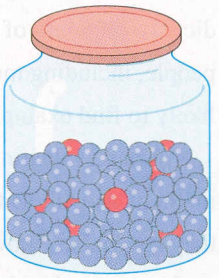
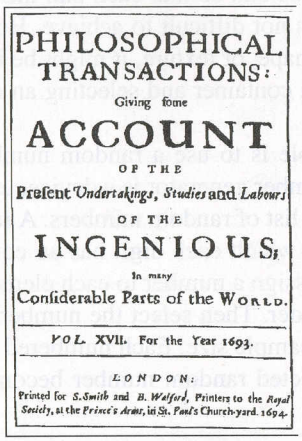


Figure 13.1

DID YOU KNOW

The Birth of Inferential Statistics


PHILOSOPHICAL TRANSACTIONS:
Giving some
ACCOUNT
OF THE
Present Undertakings, Studies and Labours
OF THE
INGENIOUS,
In many
Considerable Parts of the WORLD.
VOL. XVII. For the Year 1693.
LONDON:
Printed for S. Smith and B. Walford, Printers to the Royal Society, at the Prince's Arms, in St. Paul's Church-yard. 1694.

John Gaunt, a London merchant, is credited with being the first person to make statistical predictions, or inferences, from a set of data rather than basing the predictions simply on the laws of chance. He studied the vital statistics (births, deaths, marriages) contained in the Bills of Mortality published during the years of the Great Plague. He observed that more males were born than females and that women lived longer than men. From these observations, he made predictions about life expectancies. The keeping of mortality statistics was stimulated considerably by the growth of the insurance industry.

that the jar contains all blue marbles. If the statistician takes a larger sample, say, 15 marbles, he or she is likely to select some red marbles. At that point, the statistician may make a prediction about the contents of the jar based on the sample selected. Of course, the most accurate result would occur if every object in the jar, the entire population, were observed. However, in most statistical experiments, observing the entire population is not practical.

Statisticians use samples instead of the entire population for two reasons: (a) it is often impossible to obtain data on an entire population, and (b) sampling is less expensive because collecting the data takes less time and effort. For example, suppose that you wanted to determine the number of each species of all the fish in a lake. To do so would be almost impossible without using a sample. If you did try to obtain this information from the entire population, the cost would be astronomical. Or suppose that you wanted to test soup cans for spoilage. If every can produced by the company was opened and tested, the company wouldn't have any product left to sell. Instead of testing the entire population of soup cans, a sample is selected. The results obtained from the sample of soup cans selected are used to make conclusions about the entire population of soup cans.

Later in this chapter we will discuss statistical measures such as the *mean* and the *standard deviation*. When statisticians calculate the mean and the standard deviation of the entire population they use different symbols and formulas than when they calculate the mean and standard deviation of a sample. The following chart shows the symbols used to represent the mean and standard deviation of a sample and of a population. Note that the mean and standard deviation of a population are symbolized by Greek letters.

Measure	Sample	Population
Mean	$\bar{x}$ (read "x bar")	μ (mu)
Standard deviation	s	σ (sigma)

Unless otherwise indicated, in this book we will always assume that we are working with a sample and so we will use $\bar{x}$ and s. If you take a course in statistics, you will use all four symbols and different formulas for a sample and for a population.

Consider the task of determining the political strength of a certain candidate running in a national election. It is not possible for pollsters to ask each of the approximately 190 million eligible voters his or her preference for a candidate. Thus, pollsters must select and use a sample of the population to obtain their information. How large a sample do you think they use to make predictions about an upcoming national election? You might be surprised to learn that pollsters use only about 1600 registered voters in their national sample. How can a pollster using such a small percentage of the population make an accurate prediction?

The answer is that, when pollsters select a sample, they use sophisticated statistical techniques to obtain an unbiased sample. An ***unbiased sample*** is one that is a small replica of the entire population with regard to income, education, gender, race, religion, political affiliation, age, and so on. The procedures statisticians use to obtain unbiased samples are quite complex. The following sampling techniques will give you a brief idea of how statisticians obtain unbiased samples.

Random Sampling

If a sample is drawn in such a way that each time an item is selected each item in the population has an equal chance of being drawn, the sample is said to be a ***random sample***. When using a random sample, one combination of a specified number of items has

DID YOU KNOW

Don't Count Your Votes Until They're Cast

A classic instance of faulty sampling occurred in the 1936 presidential election. On the basis of the responses of 2,300,000 voters, selected from automobile owners and telephone subscribers, the *Literary Digest* confidently predicted that the Republican candidate, Alf Landon, would be elected. As it turned out, Franklin D. Roosevelt, the Democratic candidate, won by a large margin. The erroneous prediction occurred because the voters used in the sample were not representative of the general voting population. In 1936, telephones and automobiles were unaffordable to the average voter.

the same probability of being selected as any other combination. When all the items in the population are similar with regard to the specific characteristic we are interested in, a random sample can be expected to produce satisfactory results. For example, consider a large container holding 300 tennis balls that are identical except for color. One-third of the balls are yellow, one-third are white, and one-third are green. If the balls can be thoroughly mixed between each draw of a tennis ball so that each ball has an equally likely chance of being selected, randomness is not difficult to achieve. However, if the objects or items are not all the same size, shape, or texture, it might be impossible to obtain a random sample by reaching into a container and selecting an object.

The best procedure for selecting a random sample is to use a random number generator or a table of random numbers. A random number generator is a device, usually a calculator or computer program, that produces a list of random numbers. A random number table is a collection of random digits in which each digit has an equal chance of appearing. To select a random sample first assign a number to each element in the population. Numbers are usually assigned in order. Then select the number of random numbers needed, which is determined by the sample size. Each numbered element from the population that corresponds to a selected random number becomes part of the sample.

Systematic Sampling

When a sample is obtained by drawing every nth item on a list or production line, the sample is a ***systematic sample***. The first item should be determined by using a random number.

It is important that the list from which a systematic sample is chosen include the entire population being studied. See the Did You Know called "Don't Count Your Votes Until They're Cast." Another problem that must be avoided when this method of sampling is used is the constantly recurring characteristic. For example, on an assembly line, every 10th item could be the work of robot X. If only every 10th item is checked, the work of other robots doing the same job may not be checked and may be defective.

Cluster Sampling

A ***cluster sample*** is sometimes referred to as an *area sample* because it is frequently applied on a geographical basis. Essentially, the sampling consists of a random selection of groups of units. To select a cluster sample we divide a geographic area into sections. Then we randomly select the sections or clusters. Either each member of the selected cluster is included in the sample or a random sample of the members of each cluster is used. For example, geographically we might randomly select city blocks to use as a sample unit. Then either every member of each selected city block would be used or a random sample from each selected city block would be used. Another example would be to select x boxes of screws from a whole order, count the number of defective screws in the x boxes, and use this number to determine the expected number of defective screws in the whole order.

Stratified Sampling

When a population is divided into parts, called strata, for the purpose of drawing a sample, the procedure is known as ***stratified sampling***. Stratified sampling involves dividing the population by characteristics called *stratifying factors* such as gender,

race, religion, or income. When a population has varied characteristics, it is desirable to separate the population into classes with similar characteristics and then take a random sample from each stratum (or class). For example, we could separate the population of undergraduate college students into strata called freshmen, sophomores, juniors, and seniors.

The use of stratified sampling requires some knowledge of the population. For example, to obtain a cross section of voters in a city, we must know where various groups are located and the approximate numbers in each location.

Convenience Sampling

A ***convenience sample*** uses data that are easily or readily obtained. Occasionally, data that are conveniently obtained may be all that is available. In some cases, some information is better than no information at all. Nevertheless, convenience sampling can be extremely biased. For example, suppose that a town wants to raise taxes to build a new elementary school. The local newspaper wants to obtain the opinion of some of the residents and sends a reporter to a senior citizens center. The first 10 people who exit the building are asked if they are in favor of raising taxes to build a new school. This sample could be biased against raising taxes for the new school. Most senior citizens would not have school-age children and may not be interested in paying increased taxes to build a new school. Although a convenience sample may be very easy to select, one must be very cautious when using the results obtained from this method.

EXAMPLE 1 *Identifying Sampling Techniques*

Identify the sampling technique used to obtain a sample in the following. Explain your answer.

a) Every 20th calculator coming off an assembly line is checked for defects.

b) A $50 gift certificate is given away at the Annual Bankers Convention. Tickets are placed in a bin and the tickets are mixed up. Then the winning ticket is selected by a blindfolded person.

c) Children in a large city are classified based on the neighborhood school they attend. A random sample of five schools is selected. All the children from each selected school are included in the sample.

d) The first 30 people entering an opera house are asked if they support an increase in funding for the arts.

e) Students at Portland State University are classified according to their major. Then a random sample of 15 students from each major is selected.

SOLUTION:

a) Systematic sampling. The sample is obtained by drawing every nth item. In this example, every 20th item on an assembly line is selected.

b) Random sampling. Every ticket has an equal chance of being selected.

c) Cluster sampling. A random sample of geographic areas is selected.

d) Convenience sampling. The sample is selected by picking people that are easily obtained.

e) Stratified sampling. The students are divided into strata based on their majors. Then random samples are selected from each strata. ▲

SECTION 13.1 EXERCISES

Concept/Writing Exercises

1. Define *statistics* in your own words.
2. Explain the difference between descriptive and inferential statistics.
3. When you hear the word *statistics,* what specific words or ideas come to mind?
4. Attempt to list at least two professions in which no aspect of statistics is used.
5. Name five areas other than those mentioned in this section in which statistics is used.
6. Explain the difference between probability and statistics.
7. **a)** What is a population?
   b) What is a sample?
8. **a)** What is a systematic sample?
   b) How might a systematic sample be selected?
9. **a)** What is a random sample?
   b) How might a random sample be selected?
10. **a)** What is a cluster sample?
    b) How might a cluster sample be selected?
11. **a)** What is a stratified sample?
    b) How might a stratified sample be selected?
12. **a)** What is a convenience sample?
    b) How might a convenience sample be selected?
13. What is an unbiased sample?
14. ***Family Size*** The principal of an elementary school wishes to determine the "average" family size of the children who attend the school. To obtain a sample, the principal visits each room and selects the four students closest to each corner of the room. The principal asks each of these students how many people are in his or her family.
    a) Will this technique result in an unbiased sample? Explain your answer.
    b) If the sample is biased, will the average be greater than or less than the true family size? Explain.

Practice the Skills

Sampling Techniques *In Exercises 15–24, identify the sampling technique used to obtain a sample. Explain your answer.*

15. A group of people are classified according to age and then random samples of people from each group are taken.
16. Every 15th CD player coming off an assembly line is checked for defects.
17. A state is divided into regions using zip codes. A random sample of 20 zip code areas is selected.
18. A door prize is given away at a teachers' convention. Tickets are placed in a bin and the tickets are mixed up. Then a ticket is selected by a blindfolded person.
19. Every 17th person in line to buy tickets for a rock concert is asked his or her age.
20. The businesses in Iowa City are grouped according to type: medical, service, retail, manufacturing, financial, construction, restaurant, hotel, tourism, and other. A random sample of 10 businesses from each type is selected.
21. The first 25 students leaving the cafeteria are asked how much money they spent on textbooks for the semester.
22. The Food and Drug Administration randomly selects five stores from each of four randomly selected sections of a large city and checks food items for freshness. These stores are used as a representative sample of the entire city.
23. Bingo balls in a bin are shaken and then balls are selected from the bin.

24. The Student Senate at the University of New Orleans is electing a new president. The first 25 people leaving the library are asked for whom they will vote.

Challenge Problems/Group Activities

25. **a)** ***Random Sampling*** Select a topic and population of interest to which a random sampling technique can be applied to obtain data.
    b) Explain how you or your group would obtain a random sample for your population of interest.
    c) Actually obtain the sample by the procedure stated in part (b).
26. ***Data from Questionnaire*** Some subscribers of *Consumer Reports* respond to an annual questionnaire regarding their satisfaction with new appliances, cars, and other items. The information obtained from these questionnaires is then used as a sample from which frequency of repairs and other ratings are made by the magazine. Are the data obtained from these returned questionnaires representative of the entire population or are they biased? Explain your answer.

Recreational Mathematics

27. Statistically speaking, what is the most dangerous job in the United States?

28. Refer to the Did You Know on page 700. Select a random sample of 30 people and see how many of the 30 people have the same birthday. (*Hint:* The probability of at least 2 of the 30 sharing the same birthday is greater than 0.5).

Internet/Research Problem

29. We have briefly introduced sampling techniques. Using statistics books and Internet websites as references, select one type of sampling technique (it may be one that we have not discussed in this section) and write a report on how statisticians obtain that type of sample. Also indicate when that type of sampling technique may be preferred. List two examples of when the sampling technique may be used.

13.2 THE MISUSES OF STATISTICS

Statistics, when used properly, is a valuable tool to society. However, many individuals, businesses, and advertising firms misuse statistics to their own advantage. You should examine statistical statements very carefully before accepting them as fact. Two questions you should ask yourself are: Was the sample used to gather the statistical data unbiased and of sufficient size? Is the statistical statement ambiguous; that is, can it be interpreted in more than one way?

Let's examine two advertisements. "Four out of five dentists recommend sugarless gum for their patients who chew gum." In this advertisement, we do not know the sample size and the number of times the experiment was performed to obtain the desired results. The advertisement does not mention that possibly only 1 of 100 dentists recommended gum at all.

In a golf ball commercial, a "type A" ball is hit, and a second ball is hit in the same manner. The type A ball travels farther. We are supposed to conclude that the type A is the better ball. The advertisement does not mention the number of times the experiment was previously performed or the results of the earlier experiments. Possible sources of bias include (1) wind speed and direction, (2) that no two swings are identical, and (3) that the ball may land on a rough or smooth surface.

Vague or ambiguous words also lead to statistical misuses or misinterpretations. The word *average* is one such culprit. There are at least four different "averages," some of which are discussed in Section 13.5. Each is calculated differently, and each may have a different value for the same sample. During contract negotiations, it is not uncommon for an employer to state publicly that the average salary of its employees is $35,000, whereas the employees' union states that the average is $30,000. Who is lying? Actually, both sides may be telling the truth. Each side will use the average that best suits its needs to present its case. Advertisers also use the average that most enhances their products. Consumers often misinterpret this average as the one with which they are most familiar.

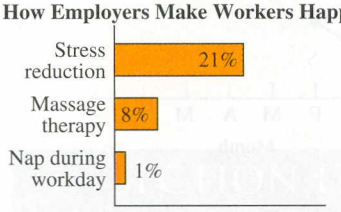
Another vague word is *largest*. For example, ABC claims that it is the largest department store in the United States. Does that mean largest profit, largest sales, largest building, largest staff, largest acreage, or largest number of outlets?

Still another deceptive technique used in advertising is to state a claim from which the public may draw irrelevant conclusions. For example, a disinfectant manufacturer claims that its product killed 40,760 germs in a laboratory in 5 seconds. "To prevent colds, use disinfectant A." It may well be that the germs killed in the laboratory were not related to any type of cold germ. In another example, company C claims that its paper towels are heavier than its competition's towels. Therefore, they will hold more water. Is weight a measure of absorbency? A rock is heavier than a sponge, yet a sponge is more absorbent.

DID YOU KNOW

Creative Displays

How Employers Make Workers Happy



Visual graphics are often used to "dress up" what might otherwise be considered boring statistics. Although visually appealing, such creative displays of numerical data can be misleading. The graph above, shows the percentage of employers that offer "perks" such as stress reduction, massage therapy, or a nap during the workday to make workers happy. This graph is misleading because the lengths of the bars are not proportional to one another as they should be to accurately reflect the percent of employers offering each of the named perks. For example, the bar for massage therapy should be eight times as long as the bar for nap during workday instead of being approximately four times as long, as the graph shows.

An insurance advertisement claims that in Duluth, Minnesota, 212 people switched to insurance company Z. One may conclude that this company is offering something special to attract these people. What may have been omitted from the advertisement is that 415 people in Duluth, Minnesota, dropped insurance company Z during the same period.

A foreign car manufacturer claims that 9 of every 10 of a popular model car it sold in the United States during the previous 10 years were still on the road. From this statement the public is to conclude that this foreign car is well manufactured and would last for many years. The commercial neglects to state that this model has been selling in the United States for only a few years. The manufacturer could just as well have stated that 9 of every 10 of these cars sold in the United States in the previous 100 years were still on the road.

Charts and graphs can also be misleading or deceptive. In Fig. 13.2, the two graphs show the performance of two stocks over a 6-month period. Based on the graphs, which stock would you purchase? Actually, the two graphs present identical information; the only difference is that the vertical scale of the graph for stock B has been exaggerated.

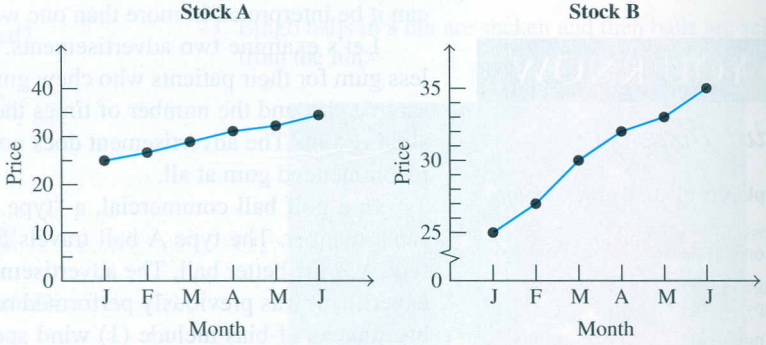


Figure 13.2

The two graphs in Fig. 13.3 show the same change. However, the graph in part (a) appears to show a greater increase than the graph in part (b), again because of a different scale.

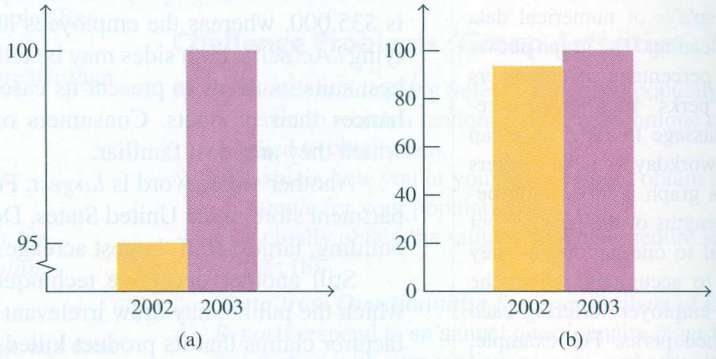


Figure 13.3

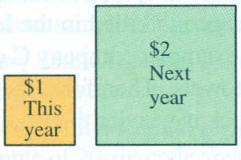


Figure 13.4

Consider a claim that if you invest $1, by next year you will have $2. This type of claim is sometimes misrepresented, as in Fig. 13.4. Actually, your investment has only doubled, but the area of the square on the right is four times that of the square on the left. By expressing the amounts as cubes (Fig. 13.5), you increase the volume eightfold.

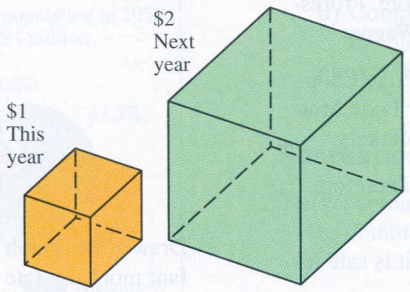


Figure 13.5

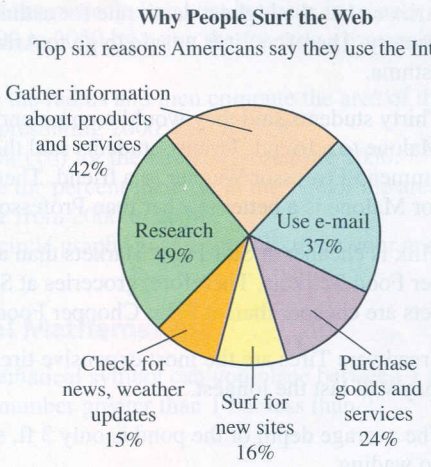


Figure 13.6

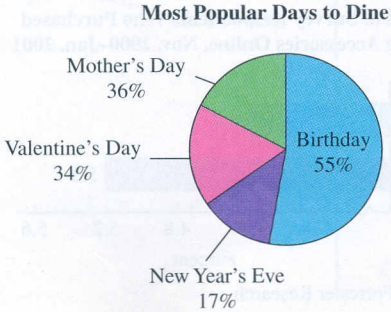
The graph in Fig. 13.6 is an example of a circle graph. We will discuss how to construct circle graphs in Section 13.4. In a circle graph, the total circle represents 100%. Therefore, the sum of the parts should add up to 100%. This graph is misleading since the sum of its parts is 183%. A graph other than a circle graph should have been used to display the top six reasons Americans say they use the Internet.

Despite the examples presented in this section, you should not be left with the impression that statistics is used solely for the purpose of misleading or cheating the consumer. As stated earlier, there are many important and necessary uses of statistics. Most statistical reports are accurate and useful. You should realize, however, the importance of being an aware consumer.

SECTION 13.2 EXERCISES

Concept/Writing Exercises

1. Find five advertisements or commercials that may be statistically misleading. Explain why each may be misleading.
2. A sample of 300 people leaving a restaurant was asked the following question: "On which of the following special occasions are you likely to dine out: your birthday, Mother's Day, Valentine's Day, or New Year's Eve?" The following circle graph shows the percent of responses for each of the above special occasions. Is the graph misleading? Explain.



Practice the Skills

Misinterpretations of Statistics *In Exercises 3–16, discuss the statement and tell what possible misuses or misinterpretations may exist.*

3. In 2003, there were more car thefts in Baltimore, Maryland, than in Reno, Nevada. Therefore, you are more likely to have your car stolen in Baltimore than in Reno.
4. There are more empty spaces in the parking lot of Mama Mia's Italian restaurant than at Shanghi Chinese restaurant. Therefore, more people prefer Chinese food than Italian food.
5. Healthy Snacks cookies are fat free. So eat as many as you like and you will not gain weight.
6. Morgan's is the largest department store in New York. So shop at Morgan's and save money.
7. Most accidents occur on Saturday night. This means that people do not drive carefully on Saturday night.
8. Eighty percent of all automobile accidents occur within 10 miles of the driver's home. Therefore, it is safer to take long trips.

9. Arizona has the highest death rate for asthma in the country. Therefore, it is unsafe to go to Arizona if you have asthma.

10. Thirty students said they would recommend Professor Malone to a friend. Twenty students said they would recommend Professor Wagner to a friend. Therefore, Professor Malone is a better teacher than Professor Wagner.

11. Milk is cheaper at Star Food Markets than at Price Chopper Food Markets. Therefore, groceries at Star Food Markets are cheaper than at Price Chopper Food Markets.

12. Treadware Tires are the most expensive tires. Therefore, they will last the longest.

13. The average depth of the pond is only 3 ft, so it is safe to go wading.

14. More men than women are involved in automobile accidents. Therefore, women are better drivers.

15. At West High School, half the students are below average in mathematics. Therefore, the school should receive more federal aid to raise student scores.

16. More women than men applied for admission to the 2003 class of Mesa Community College. Therefore, in 2003, more women than men will attend Mesa Community College.

17. ***Hospital Care Expenditures*** The following table shows the percent of national health expenditures spent on hospital care for selected years.

Year	Percent
1985	39.1
1990	36.5
1995	34.7
1997	33.6
1998	33.0
1999	32.3
2000	31.7

Source: National Center for Health Statistics, U.S. Department of Health and Human Services.

Draw a line graph that makes the decrease in the percent of national health expenditures spent on hospital care from 1985 through 2000 appear to be
a) small.
b) large.

18. ***Infant Mortality Rate*** The table above and to the right shows the United States infant mortality rate, per 1000 births from 1994 to 2000.

Year	Rate
1994	8.0
1995	7.6
1996	7.3
1997	7.2
1998	7.2
1999	7.1
2000	6.9

Source: National Center for Health Statistics, U.S. Department of Health and Human Services.

Draw a line graph that makes the decrease in the U.S. infant mortality rate from 1994 through 2000 appear to be
a) small.
b) large.

First Marriage *In Exercises 19 and 20, use the following table.*

Median Age at First Marriage

Male		Female	
Year	**Age**	**Year**	**Age**
1970	23.2	1970	20.8
1980	24.7	1980	22.0
1990	26.1	1990	23.9
2000	26.9	2000	25.1

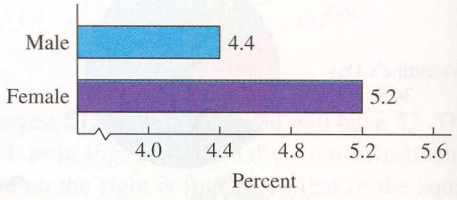
Source: U.S. Census Bureau.

19. **a)** Draw a bar graph that appears to show a small increase in the median age at first marriage for males.
b) Draw a bar graph that appears to show a large increase in the median age at first marriage for males.

20. **a)** Draw a bar graph that appears to show a small increase in the median age at first marriage for females.
b) Draw a bar graph that appears to show a large increase in the median age at first marriage for females.

21. ***Online Purchasing*** The following graph shows the percent of males and the percent of females surveyed that purchased clothing accessories online during the months from November 2000 to January 2001.
a) Draw a bar graph that shows the entire scale from 0 to 6.
b) Does the new graph give a different impression? Explain.

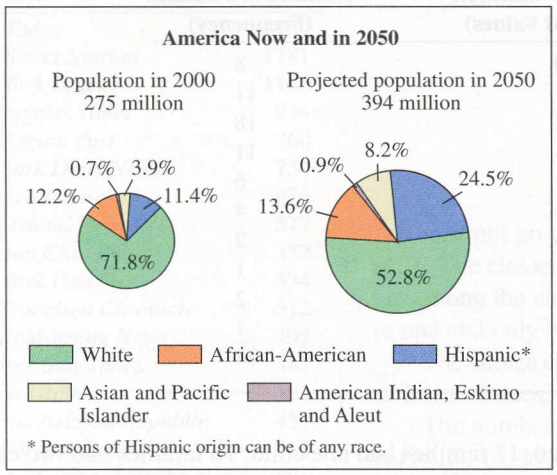
Percent of Survey Respondents Who Purchased Clothing Accessories Online, Nov. 2000–Jan. 2001



Source: Forrester Research

Challenge Problem/Group Activity

22. Consider the following graph, which shows the U.S. population in 2000 and the projected U.S. population in 2050.



Source: U.S. Census Bureau

a) Compute the projected percent increase in population from 2000 to 2050 by using the formula given on page 595.
b) Measure the radius and then compute the area of the circle representing 2000. Use $A = \pi r^2$.
c) Repeat part (b) for the circle representing 2050.
d) Compute the percent increase in the size of the area of the circle from 2000 to 2050.
e) Are the circle graphs misleading? Explain your answer.

Recreational Mathematics

23. What mathematical symbol can you place between 1 and 2 to obtain a number greater than 1 but less than 2?

Internet/Research Activity

24. Read the book *How to Lie with Statistics* by Darrell Huff and write a book report on it. Select three illustrations from the book that show how people manipulate statistics.

13.3 FREQUENCY DISTRIBUTIONS

"Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write."
H. G. Wells

It is not uncommon for statisticians and others to have to analyze thousands of pieces of data. A ***piece of data*** is a single response to an experiment. When the amount of data is large, it is usually advantageous to construct a frequency distribution. A ***frequency distribution*** is a listing of the observed values and the corresponding frequency of occurrence of each value.

EXAMPLE 1 *Frequency Distribution*

The number of children per family is recorded for 64 families surveyed. Construct a frequency distribution of the following data:

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

DID YOU KNOW

Can You Count the F's?

Statistical errors often result from careless observations. To see how such errors can occur, consider the statement below. How many F's do you count in the statement? You can find the answer in the answer section (Section 13.3, Exercise 3.2).

FINISHED FILES ARE THE RESULT OF YEARS OF SCIENTIFIC STUDY COMBINED WITH THE EXPERIENCE OF YEARS.

SOLUTION: Listing the number of children (observed values) and the number of families (frequency) gives the following frequency distribution.

Number of Children (Observed Values)	Number of Families (Frequency)
0	8
1	11
2	18
3	11
4	6
5	4
6	2
7	1
8	2
9	1
	64

Eight families had no children, 11 families had one child, 18 families had two children, and so on. Note that the sum of the frequencies is equal to the original number of pieces of data, 64. ▲

Often data are grouped in classes to provide information about the distribution that would be difficult to observe if the data were ungrouped. Graphs called *histograms* and *frequency polygons* can be made of grouped data, as will be explained in Section 13.4. These graphs also provide a great deal of useful information.

When data are grouped in classes, certain rules should be followed.

Rules for Data Grouped by Classes

1. The classes should be of the same "width."
2. The classes should not overlap.
3. Each piece of data should belong to only one class.

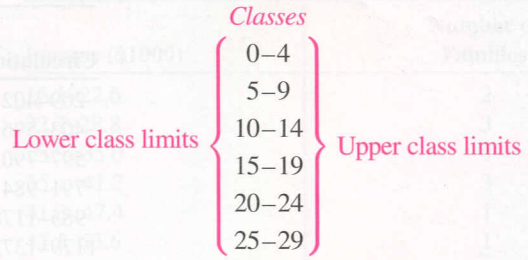
In addition, it is often suggested that a frequency distribution should be constructed with 5 to 12 classes. If there are too few or too many classes, the distribution may become difficult to interpret. For example, if you use fewer than 5 classes, you risk losing too much information. If you use more than 12 classes, you may gain more detail but you risk losing clarity. Let the spread of the data be a guide in deciding the number of classes to use.

To understand these rules, let's consider a set of observed values that go from a low of 0 to a high of 26. Let's assume that the first class is arbitrarily selected to go from 0 through 4. Thus, any of the data with values of 0, 1, 2, 3, 4 would belong in this class. We say that the ***class width*** is 5, since there are five integral values that belong to the class. This first class ended with 4, so the second class must start with 5. If this class is to have a width of 5, at what value must it end? The answer is 9 (5, 6, 7, 8, 9). The second class is 5–9. Continuing in the same manner, we obtain the following set of classes.

TABLE 13.1

Newspaper	Circulation (thousands)
USA Today	2150
Wall Street Journal	1781
New York Times	1109
Los Angeles Times	944
Washington Post	760
New York Daily News	734
Chicago Tribune	676
Long Island Newsday	577
Houston Chronicle	552
New York Post	534
San Francisco Chronicle	512
Dallas Morning News	495
Chicago Sun-Times	481
Boston Globe	471
Phoenix Arizona Republic	451
Newark Star-Ledger	411
Atlanta Journal-Constitution	396
Detroit Free Press	371
Philadelphia Inquirer	365
Cleveland Plain Dealer	360
San Diego Union-Tribune	352
Portland Oregonian	351
Minneapolis Star Tribune	340
St. Petersburg Times	332
Orange County Register	324
Miami Herald	318
Denver Rocky Mountain News	310
Baltimore Sun	306
Denver Post	306
St. Louis Post-Dispatch	291
Sacramento Bee	286
Investor's Business Daily	281
San Jose Mercury News	269
Kansas City Star	260
Boston Herald	259
Milwaukee Journal Sentinel	255
Orlando Sentinel	255
New Orleans Times-Picayune	255
Indianapolis Star	252
Fort Lauderdale Sun-Sentinel	252
Columbus Dispatch	244
Detroit News	243
Pittsburgh Post-Gazette	242
Charlotte Observer	235
Louisville Courier-Journal	222
Seattle Times	219
Buffalo News	219
Fort Worth Star-Telegram	214
Tampa Tribune	213
San Antonio Express-News	209

Source: *2002 Editor & Publisher International Yearbook.*



We need not go beyond the 25–29 class because the largest value we are considering is 26. The classes meet our three criteria: They have the same width, there is no overlap among the classes, and each of the values from a low of 0 to a high of 26 belongs to one and only one class.

The choice of the first class, 0–4, was arbitrary. If we wanted to have more classes or fewer classes, we would make the class widths smaller or larger, respectively.

The numbers 0, 5, 10, 15, 20, 25 are called the ***lower class limits***, and the numbers 4, 9, 14, 19, 24, 29 are called the ***upper class limits***. Each class has a width of 5. Note that the class width, 5, can be obtained by subtracting the first lower class limit from the second lower class limit: $5 - 0 = 5$. The difference between any two consecutive lower class or upper class limits is also 5.

EXAMPLE 2 *A Frequency Distribution of Daily Newspaper Circulation*

Table 13.1 in the margin shows the circulation for the 50 leading U.S. daily newspapers, in 2001. The circulation is rounded to the nearest thousand. Construct a frequency distribution of the data, letting the first class be 209–402.

SOLUTION: Fifty pieces of data are given in *descending order* from highest to lowest. We are given that the first class is 209–402. The second class must therefore start at 403. To find the class width we subtract 209 (the lower class limit of the first class) from 403 (the lower class limit of the second class) to obtain a class width of 194. The upper class limit of the second class is found by adding the class width, 194, to the upper class limit of the first class, 402. Therefore, the upper class limit of the second class is $402 + 194 = 596$. Thus,

$$209 - 402 = \text{first class}$$
$$403 - 596 = \text{second class}$$

The other classes are found using a similar technique. The other classes are 597–790, 791–984, 985–1178, 1179–1372, 1373–1566, 1567–1760, 1761–1954, 1955–2148, and 2149–2342. Since the highest value in the data is 2150, there is no need to go any further. Note that each two consecutive lower class limits differ by 194, as do each two consecutive upper class limits. There are 34 pieces of data in the 209–402 class. There are 9 pieces of data in the 403–596 class, 3 in the 597–790 class, 1 in the 791–984 class, 1 in the 985–1178 class, 0 in the 1179–1372 class, 0 in the 1373–1566 class, 0 in the 1567–1760 class, 1 in the 1761–1954 class, 0 in the 1955–2148 class, and 1 in the 2149–2342 class. The complete frequency distribution of the 11 classes is given on page 764. The number of newspapers totals 50, so we have included each piece of data.

Circulation	Number of Newspapers
209–402	34
403–596	9
597–790	3
791–984	1
985–1178	1
1179–1372	0
1373–1566	0
1567–1760	0
1761–1954	1
1955–2148	0
2149–2342	1
	50

The ***modal class*** of a frequency distribution is the class with the greatest frequency. In Example 2, the modal class is 209–402. The ***midpoint of a class***, also called the ***class mark***, is found by adding the lower and upper class limits and dividing the sum by 2. The midpoint of the first class in Example 2 is

$$\frac{209 + 402}{2} = \frac{611}{2} = 305.5$$

Note that the difference between successive class marks is the class width. The class mark of the second class can therefore be obtained by adding the class width, 194, to the class mark of the first class, 305.5. The sum is $305.5 + 194 = 499.5$. Note that $\frac{403 + 596}{2} = \frac{999}{2} = 499.5$ which checks with the class mark obtained by adding the class width to the first class mark.

EXAMPLE 3 *A Frequency Distribution of Family Income*

The following set of data represents the family income (in thousands of dollars, rounded to the nearest hundred) of 15 randomly selected families.

31.5	16.8	30.8	29.7	25.9
50.2	37.4	29.6	38.7	33.8
20.5	25.3	24.8	41.3	35.7

Construct a frequency distribution with a first class of 16.5–22.6.

SOLUTION: First rearrange the data from lowest to highest so that the data will be easier to categorize.

16.8	25.3	29.7	33.8	38.7
20.5	25.9	30.8	35.7	41.3
24.8	29.6	31.5	37.4	50.2

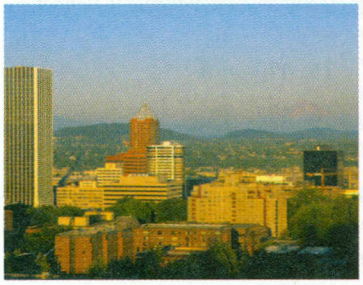
The first class goes from 16.5 to 22.6. Since the data are in tenths, the class limits will also be given in tenths. The first class ends with 22.6; therefore, the second class must start with 22.7. The class width of the first class is $22.7 - 16.5$, or 6.2. The upper class limit of the second class must therefore be $22.6 + 6.2$, or 28.8. The frequency distribution is as follows.

Income ($1000)	Number of Families
16.5–22.6	2
22.7–28.8	3
28.9–35.0	5
35.1–41.2	3
41.3–47.4	1
47.5–53.6	1
	15

Note in Example 3 that the class width is 6.2, the modal class is 28.9 − 35.0, and the class mark of the first class is (16.5 + 22.6)/2, or 19.55.

DID YOU KNOW

Cyberspace Is the Place to Be



Portland, Oregon, the "most wired" city.

The Internet, first developed as a communications tool for the government and for research universities, is now being used for everything from research to entertainment to shopping to chatting with friends. Internet usage continues to increase dramatically. In 2001, 56.5% of households in the United States had computers, and 50.5% of households in the United States were online. Thirty-nine percent of Internet users made online purchases and 35% of Internet users searched for health information. Consumers are also using the Internet to provide themselves with automotive information prior to purchasing a new vehicle. According to a study done by J.D. Power and Associates, 62% of all new vehicle buyers in 2001 used the Internet for automotive information before purchasing their vehicle. Internet usage was growing at a rate of 2 million new users per month in 2001. According to Nielsen/NetRatings, as of March 2001, Portland, Oregon, was the overall "most wired" city in the United States, with almost 70% of households having Internet access from a personal computer at home.

SECTION 13.3 EXERCISES

Concept/Writing Exercises

1. What is a frequency distribution?
2. How can a class width be determined using class limits?
3. Suppose that the first class of a frequency distribution is 9–15.
   a) What is the width of this class?
   b) What is the second class?
   c) What is the lower class limit of the second class?
   d) What is the upper class limit of the second class?
4. Repeat Exercise 3 for a frequency distribution whose first class is 12–20.
5. What is the modal class of a frequency distribution?
6. What is another name for the midpoint of a class? How is the midpoint of a class determined?

Practice the Skills/Problem Solving

In Exercises 7 and 8, use the frequency distribution to determine

a) *the total number of observations.* - sum of frequencies
b) *the width of each class.*
c) *the midpoint of the second class.*
d) *the modal class (or classes).*
e) *the class limits of the next class if an additional class were to be added.*

7.

Class	Frequency
9–15	3
16–22	6
23–29	1
30–36	0
37–43	3
44–50	5

8.

Class	Frequency
40–49	7
50–59	5
60–69	3
70–79	2
80–89	7
90–99	1

9. ***Sales*** A car dealership is interested in the number of cars sold daily. A sample is taken over 40 days to obtain the following data regarding the number of cars sold daily. Construct a frequency distribution, letting each class have a width of 1 (as in Example 1).

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

10. ***Park Visits by Families*** The town of Brighton is planning to improve the local park. The responses of 32 families who were asked how many times per year they visit the park are shown below. Construct a frequency distribution letting each class have a width of 1.

20	21	24	25	26	27	29	32
20	23	24	25	26	27	30	32
20	23	24	26	26	28	31	33
21	23	24	26	26	28	31	34

Note: No one visited the park 22 times per year. However, it is customary to include a missing value as an observed value and assign to it a frequency of 0.

IQ Scores *In Exercises 11–14, use the following data, which show the result of 50 sixth-grade I.Q. scores.*

80	89	92	95	97	100	102	106	110	120
81	89	93	95	98	100	103	108	113	120
87	90	94	97	99	100	103	108	114	122
88	91	94	97	100	100	103	108	114	128
89	92	94	97	100	101	104	109	119	135

Use this data to construct a frequency distribution with a first class of

11. 78–86.

12. 80–88.

13. 80–90.

14. 80–92.

Placement Test Scores *In Exercises 15–18, use the following data, which represent the English placement test scores of a sample of 30 students.*

559	482	490	520	514
498	472	490	523	491
480	490	562	486	491
498	543	506	539	576
508	509	499	515	501
593	512	510	577	533

Use this data to construct a frequency distribution with a first class of

15. 472–492.

16. 470–486.

17. 472–487.

18. 472–496.

Newspaper Circulation *In Exercises 19–22, use the data in Example 2 on page 763 to construct a frequency distribution with a first class (in thousands) of*

19. 209–458.

20. 205–414.

21. 209–408.

22. 209–358.

County Population *In Exercises 23–26, use the following data, which represent the 2000 population of the 25 largest counties in the United States, in millions of people (rounded to the nearest 100,000).*

9.5	2.8	2.1	1.5	1.4
5.4	2.5	1.7	1.5	1.4
3.4	2.3	1.7	1.5	1.4
3.1	2.3	1.7	1.5	1.4
2.8	2.2	1.6	1.4	1.4

Use this data to construct a frequency distribution with a first class of

23. 1.4–2.1.

24. 1.0–2.7.

25. 1.0–2.5.

26. 1.4–2.9.

Price of Eggs *In Exercises 27–30, use the data in the following table.*

Average Price per Dozen Eggs for Selected States in 2000

State	Price ($)	State	Price ($)
AL	1.31	MT	0.46
AR	1.06	NE	0.38
CA	0.45	NH	0.86
CO	0.70	NJ	0.53
CT	0.56	NY	0.56
DE	0.67	NC	1.07
FL	0.48	OH	0.50
GA	0.87	OK	0.84
HI	0.89	OR	0.48
ID	0.61	PA	0.55
IL	0.47	RI	0.63
IN	0.52	SC	0.64
IA	0.38	SD	0.35
KS	0.39	TN	1.24
KY	0.90	TX	0.70
LA	0.81	UT	0.43
ME	0.60	VT	0.60
MD	0.60	VA	0.96
MA	0.63	WA	0.55
MI	0.42	WV	1.46
MS	1.18	WI	0.48
MO	0.52		

Source: National Agricultural Statistics Service, U.S. Dept. of Agriculture.

Construct a frequency distribution with a first class of

27. 0.35–0.44.

28. 0.35–0.45.

29. 0.35–0.54.

30. 0.35–0.48.

Recreational Mathematics

31. In what month do people take the least number of daily vitamins?

32. a) Count the number of F's in the sentence at the bottom of the Did You Know on page 762.

b) Can you explain why so many people count the number of F's incorrectly?

13.4 STATISTICAL GRAPHS

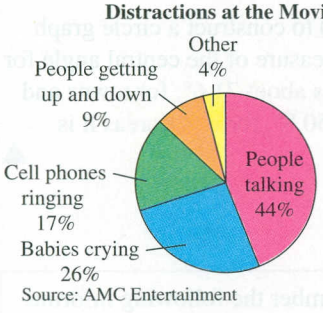


Figure 13.7

Now we will consider four types of graphs: the circle graph, the histogram, the frequency polygon, and the stem-and-leaf display.

Circle graphs (also known as pie charts) are often used to compare parts of one or more components of the whole to the whole. The circle graph in Fig. 13.7 shows what moviegoers say is the most annoying distraction during a movie. Since the total circle represents 100%, the sum of the percents of the sectors should be 100%, and it is.

In the next example, we will discuss how to construct a circle graph given a set of data.

EXAMPLE 1 *Labor Day Travel*

According to the American Automobile Association (AAA), 27.7 million Americans traveled by car during Labor Day weekend in 2001. The following table indicates the destinations of these travelers.

Destination	Number of People (millions)
Major cities	6.4
Oceans and beaches	5.5
Towns and rural areas	5.3
Mountains	3.9
Other	6.6
	27.7

Use this information to construct a circle graph illustrating the percent of people who traveled to major cities, oceans and beaches, towns and rural areas, the mountains, and other places during Labor Day weekend in 2001.

SOLUTION: Determine the measure of the corresponding central angle, as illustrated in the following table.

Destination	Number of People (millions)	Percent of Total (to the nearest tenth of a percent)	Measure of Central Angle (degrees)
Major cities	6.4	$\frac{6.4}{27.7} \times 100 = 23.1\%$	$0.231 \times 360 = 83.2°$
Oceans and beaches	5.5	$\frac{5.5}{27.7} \times 100 = 19.9\%$	$0.199 \times 360 = 71.6°$
Towns and rural areas	5.3	$\frac{5.3}{27.7} \times 100 = 19.1\%$	$0.191 \times 360 = 68.8°$
Mountains	3.9	$\frac{3.9}{27.7} \times 100 = 14.1\%$	$0.141 \times 360 = 50.8°$
Other	6.6	$\frac{6.6}{27.7} \times 100 = 23.8\%$	$0.238 \times 360 = 85.7°$
Total	27.7	100.0%	360.1° *

*Due to rounding we get 360.1°, not exactly 360°. If the measure of the central angle were rounded to hundredths, the sum would be exactly 360°.

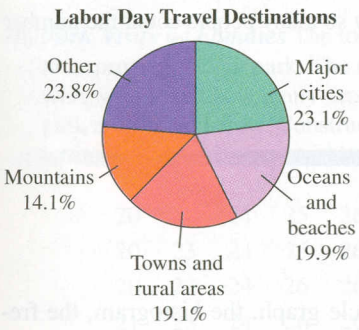


Figure 13.8

Now use a protractor (See Section 9.1, page 479.) to construct a circle graph and label it properly, as illustrated in Fig. 13.8. The measure of the central angle for major cities is about 83.2°, for oceans and beaches it is about 71.6°, for towns and rural areas it is about 68.8°, for mountains it is about 50.8°, for other areas it is about 85.7°. ▲

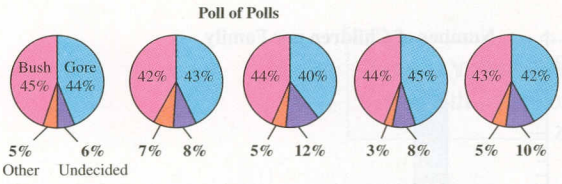
TIMELY TIP When constructing circle graphs remember the following information:

If you have to round your percent in the percent of total column, the sum of the percents may not be exactly 100%. Due to rounding the percents, the sum may be either slightly below 100% or slightly above 100%.

When calculating the measure of a central angle, if you have to round the central angle measure, the sum of the angles may not be exactly 360°. Due to rounding measurements, the sum of the angles may be slightly above 360° or slightly below 360°.

DID YOU KNOW

Polls



	Pew Research Center	CBS News	Voter.com	CNN/USA Today/Gallup	Reuters/ MSNBC
Adults polled	722	687	1,000	690	1,200
Date of poll	Oct. 4–8	Oct. 6–9	Oct. 8–11	Oct. 9–11	Oct. 10–12
Margin of error	±3 percentage points	±3 percentage points	±3.1 percentage points	±4 percentage points	±3 percentage points

During every major election we are told the results of many polls. The following circle graphs, shown in the October 23, 2000, issue of *U.S. News and World Report*, show the results of various polls taken less than a month before the 2000 U.S. presidential election. The polls were repeated at various times closer to election time. Notice that no two polling services had exactly the same results. Also notice that the margin of error varied between ±3 points to ±4 points. The margin of error is determined by the size of the sample used and by other items. Many people believe that exit polls, taken right after people vote, should not be allowed to be broadcast until after the election results are finalized, because these polls may influence the people who have not yet voted. For example, since the people on the East Coast vote about 3 hours earlier than people on the West Coast, polls showing voters' preferences in eastern states could affect voters' choices in western states.

Histograms and frequency polygons are statistical graphs used to illustrate frequency distributions. A ***histogram*** is a graph with observed values on its horizontal scale and frequencies on its vertical scale. A bar is constructed above each observed value (or class when classes are used), indicating the frequency of that value. The horizontal scale need not start at zero, and the calibrations on the horizontal and vertical scales do not have to be the same. The vertical scale must start at zero. To accommodate large frequencies on the vertical scale, it may be necessary to break the scale. Because histograms and other bar graphs are easy to interpret visually, they are used a great deal in newspapers and magazines.

EXAMPLE 2 *Construct a Histogram*

The frequency distribution developed in Example 1, Section 13.3, is repeated here. Construct a histogram of this frequency distribution.

Number of Children (Observed Values)	Number of Families (Frequency)
0	8
1	11
2	18
3	11
4	6
5	4
6	2
7	1
8	2
9	1

SOLUTION: The vertical scale must extend at least to the number 18, since that is the greatest recorded frequency (see Fig. 13.9 on page 770). The horizontal scale must include the numbers 0–9, the number of children observed. Eight families have no children. We indicate this by constructing a bar above the number 0, centered at 0, on the horizontal scale extended up to 8 on the vertical scale. Eleven families have one child, so we construct a bar extending to 11 above the number 1,

centered at 1, on the horizontal scale. We continue this procedure for each observed value. Both the horizontal and vertical scales should be labeled, the bars should be the same width and centered at the observed value, and the histogram should have a title. In a histogram, the bars should always touch.

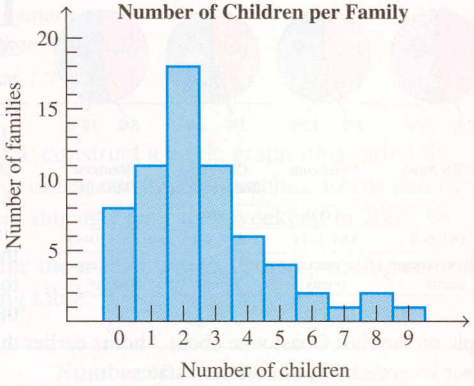


Figure 13.9

Frequency polygons are line graphs with scales the same as those of the histogram; that is, the horizontal scale indicates observed values and the vertical scale indicates frequency. To construct a frequency polygon, place a dot at the corresponding frequency above each of the observed values. Then connect the dots with straight-line segments. When constructing frequency polygons, always put in two additional class marks, one at the lower end and one at the upper end on the horizontal scale (values for these added class marks are not needed on the frequency polygon). Since the frequency at these added class marks is 0, the end points of the frequency polygon will always be on the horizontal scale.

EXAMPLE 3 *Construct a Frequency Polygon*

Construct a frequency polygon of the frequency distribution in Example 2.

SOLUTION: Since eight families have no children, place a mark above the 0 at 8 on the vertical scale, as shown in Fig. 13.10. Because there are 11 families with one

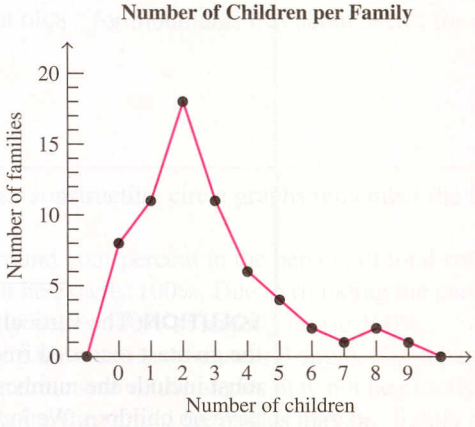


Figure 13.10

child, place a mark above the 1 on the horizontal scale at the 11 on the vertical scale, and so on. Connect the dots with straight-line segments, and bring the end points of the graph down to the horizontal scale, as shown. ▲

TIMELY TIP When constructing a histogram or frequency polygon, be sure to label both scales of the graph.

EXAMPLE 4 *Gas Mileage*

The frequency distribution of average gas mileage for selected 2003 automobiles is listed in Table 13.2. Construct a histogram and then construct a frequency polygon on the histogram.

TABLE 13.2

Mileage (mpg)	Number of Cars
10–14	5
15–19	31
20–24	36
25–29	8
30–34	5
35–39	4
40–44	1

SOLUTION: The histogram can be constructed with either class limits or class marks (class midpoints) on the horizontal scale. Frequency polygons are constructed with class marks on the horizontal scale. Since we will construct a frequency polygon on the histogram, we will use class marks. Recall that class marks are found by adding the lower class limit and upper class limit and dividing the sum by 2. For the first class, the class mark is $(10 + 14)/2$, or 12. Since the class widths are five units, the class marks will also differ by five units (see Fig. 13.11).

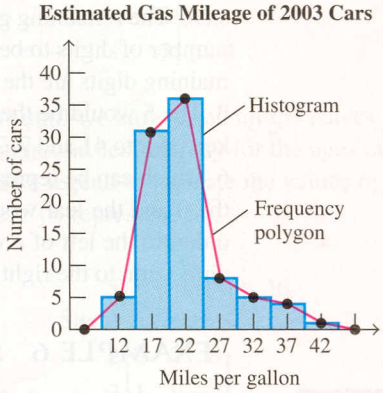


Figure 13.11 ▲

EXAMPLE 5 *Carry-on Luggage Weights*

The histogram in Fig. 13.12 on page 772 shows the weights of selected pieces of carry-on luggage at an airport. Construct the frequency distribution from the histogram in Fig. 13.12.

TABLE 13.3

Weight (pounds)	Number of Pieces of Luggage
1–5	8
6–10	10
11–15	7
16–20	5
21–25	6
26–30	3
31–35	1
36–40	2

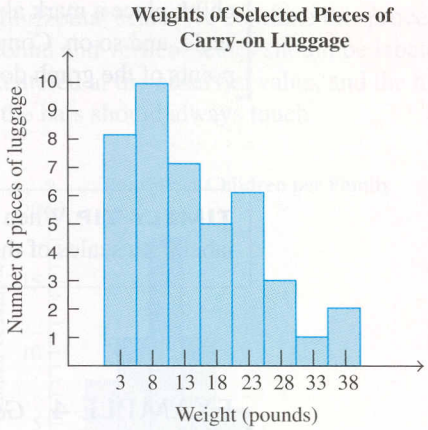


Figure 13.12

SOLUTION: There are five units between class midpoints, so each class width must also be five units. Since three is the midpoint of the first class, there must be two units below and two above it. The first class must be 1–5. The second class must therefore be 6–10. The frequency distribution is given in Table 13.3. ▲

Frequency distributions and histograms provide very useful tools to organize and summarize data. However, if the data are grouped, we cannot identify specific data values in a frequency distribution and in a histogram. For example, in Example 5, we know that there are eight pieces of luggage in the class of 1 to 5 pounds, but we don't know the specific weights of those eight pieces of luggage.

A *stem-and-leaf display* is a tool that organizes and groups the data while allowing us to see the actual values that make up the data. To construct a stem-and-leaf display each value is represented with two different groups. The left group of digits is called the *stem*. The remaining group of digits on the right is called the *leaf*. There is no rule for the number of digits to be included in the stem. Usually the units digit is the leaf and the remaining digits are the stem. For example, the number 53 would be broken up into 5 and 3. The 5 would be the stem and the 3 would be the leaf. The number 417 would be broken up into 41 and 7. The 41 would be the stem and the 7 would be the leaf. The number 6, which can be represented as 06, would be broken up into 0 and 6. The stem would be the 0 and the leaf would be the 6. With a stem-and-leaf display, the stems are listed, in order, to the left of a vertical line. Then we place each leaf to the right of its corresponding stem, to the right of the vertical line.* Example 6 illustrates this procedure.

EXAMPLE 6 *Stem-and-Leaf Display*

The table below indicates the ages of a sample of 20 guests who stayed at Captain Fairfield House Bed and Breakfast. Construct a stem-and-leaf display.

29	31	39	43	56
60	62	59	58	32
47	27	50	28	71
72	44	45	44	68

*In stem-and-leaf displays, the leaves are sometimes listed from lowest digit to greatest digit, but this is not necessary.

SOLUTION: By quickly glancing at the data, we can see the ages consist of two digit numbers. Let's use the first digit, the tens digit, as our stem and the second digit, the units digit, as the leaf. For example, for an age of 62, the stem is 6, and the leaf is 2. Our values are numbers in the 20s, 30s, 40s, 50s, 60s, and 70s. Therefore, the stems will be 2, 3, 4, 5, 6, 7 as shown below.

$$\begin{array}{r|} 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \end{array}$$

Next we place each leaf on its stem. We will do so by placing the second digit of each value next to its stem, to the right of the vertical line. Our first value is 29. The 2 is the stem and the 9 is the leaf. Therefore, we place a 9 next to the stem of 2 and to the right of the vertical line.

$$\begin{array}{r|l} 2 & 9 \end{array}$$

The next value is 31. We will place a leaf of 1 next to the stem of 3.

$$\begin{array}{r|l} 2 & 9 \\ 3 & 1 \end{array}$$

The next value is 39. Therefore, we will place a leaf of 9 after the leaf of 1 that is next to the stem of 3.

$$\begin{array}{r|l} 2 & 9 \\ 3 & 1 \quad 9 \end{array}$$

We continue this process until we have listed all the leaves on the display. The diagram below shows the stem-and-leaf display for the ages of the guests. In our display, we will also include a legend to indicate the values represented by the stems and leaves. For example, 5 | 6 represents 56.

5 | 6 represents 56

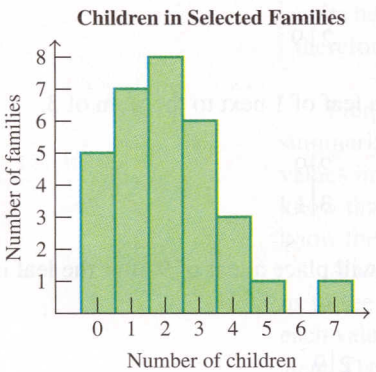
Stem	Leaves
2	9 7 8
3	1 9 2
4	3 7 4 5 4
5	6 9 8 0
6	0 2 8
7	1 2

Every piece of the original data can be seen in a stem-and-leaf display. From the above diagram, we can see that five of the guests' ages were in the 40s. Only two guests were older than 70. Note that the stem-and-leaf display gives the same visual impression as a sideways histogram.

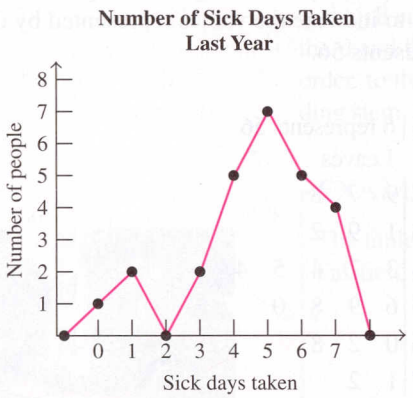
SECTION 13.4 EXERCISES

Concept/Writing Exercises

1. In your own words, explain how to construct a circle graph from a table of values.
2. **a)** What is listed on the horizontal axis of a histogram and frequency polygon?
   b) What is listed on the vertical axis of a histogram and frequency polygon?
3. In your own words, explain how to construct a frequency polygon from a set of data.
4. In your own words, explain how to construct a histogram from a set of data.
5. **a)** In your own words, explain how to construct a frequency polygon from a histogram.
   b) Construct a frequency polygon from the histogram below.



6. **a)** In your own words, explain how to construct a histogram from a frequency polygon.
   b) Construct a histogram from the frequency polygon below.



7. **a)** In your own words, explain how to construct a stem-and-leaf display.
   b) Construct a frequency distribution, letting each class have a width of 1, from the stem-and-leaf display above and to the right.

4 | 5 represents 45

Stem	Leaf
4	5 5 5 7 9
5	0 1 1

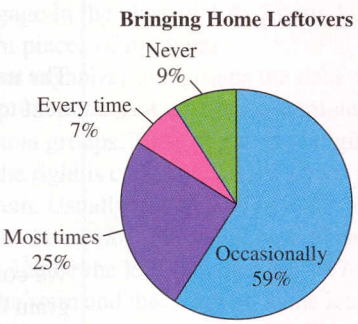
8. Construct a frequency distribution, letting each class have a width of 1, from the following stem-and-leaf display.

2 | 3 represents 23

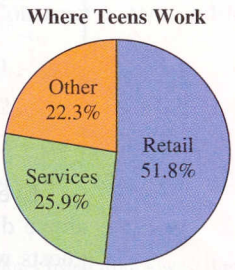
Stem	Leaf
1	7 8 7 9 6
2	3 1 2 2 5 5 4

Practice the Skills

9. *Bringing Home Leftovers* ConAgra foods, along with the American Dietetic Association, conducted a survey asking the question, "How often do you bring home leftovers from restaurants?" The following circle graph shows the percent of respondents who answered occasionally, most times, every time, or never. If 500 people were surveyed, determine the number of people in each category.



10. *Where Teens Work* The National Academy Press surveyed teens to determine where they work. The following circle graph shows the percent of 15- to 17-year-old teens surveyed who answered retail, services, or other as their type of employment. If 700 teens were surveyed, determine the number of teens working in each category. Round answers to nearest person.



Source: National Academy Press

11. ***Online Travel Websites*** A sample of 500 travelers who used the Internet to book a trip were asked which travel website they used. Their responses are given in the table below. Construct a circle graph, with sectors given in percent, which illustrates this information.

Online Travel Website	Number of Bookings
Travelocity	175
Expedia	125
Priceline	85
Other	115

12. ***Housing Permits*** A sample of 600 housing permits for new houses was randomly selected. The number of bedrooms listed in the permits was recorded, as indicated in the following table. Construct a circle graph, with sectors given in percent, which illustrates this information. Round percents to tenths.

Number of Bedrooms	Number of Permits
2	182
3	230
4	100
5 or more	88

13. ***Heights*** The frequency distribution indicates the heights of 45 male high school seniors.

Height (in.)	Number of Males
64	2
65	6
66	7
67	9
68	10
69	6
70	3
71	0
72	2

a) Construct a histogram of the frequency distribution.
b) Construct a frequency polygon of the frequency distribution.

14. ***Jazz Concert*** The frequency distribution indicates the ages of a group of 45 people attending a jazz concert.

Age	Number of People
17	2
18	5
19	7
20	8
21	0
22	10
23	5
24	8

a) Construct a histogram of the frequency distribution.
b) Construct a frequency polygon of the frequency distribution.

15. ***DVDs*** The frequency distribution indicates the number of DVDs owned by a sample of 40 people.

Number of DVDs	Number of People
6–13	4
14–21	5
22–29	10
30–37	11
38–45	6
46–53	3
54–61	1

a) Construct a histogram of the frequency distribution.
b) Construct a frequency polygon of the frequency distribution.

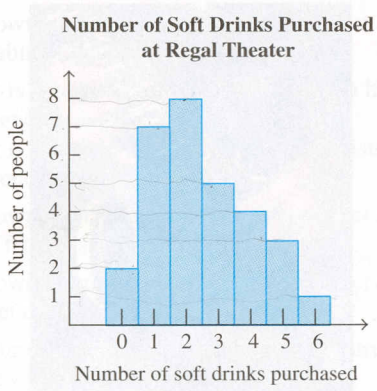
16. ***Annual Salaries*** The frequency distribution illustrates the annual salaries, in thousands of dollars, of the people in management positions at the X-Chek Corporation.

Salary (in $1000)	Number of People
30–35	4
36–41	7
42–47	8
48–53	9
54–59	7
60–65	5
66–71	3

a) Construct a histogram of the frequency distribution.
b) Construct a frequency polygon of the frequency distribution.

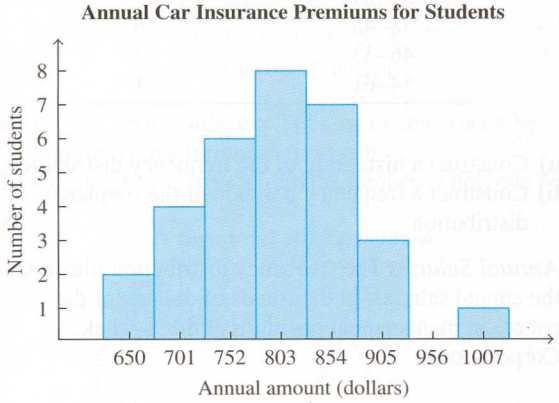
Problem Solving

17. *Number of Soft Drinks Purchased* Use the histogram below to answer the following questions.



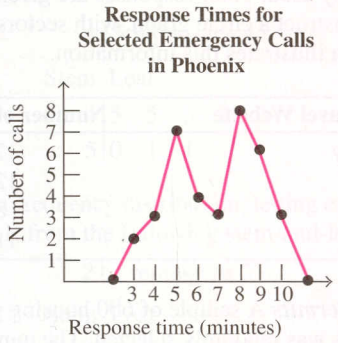
a) How many people were surveyed?
b) How many people purchased four soft drinks?
c) What is the modal class?
d) How many soft drinks were purchased?
e) Construct a frequency distribution from this histogram.

18. *Car Insurance* Use the histogram below to answer the following questions.



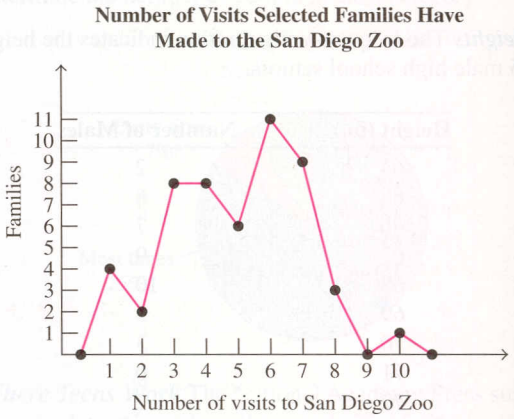
a) How many students were surveyed?
b) What are the lower and upper class limits of the first and second classes?
c) How many students have an annual car insurance premium in the class with a class mark of $752?
d) What is the class mark of the modal class?
e) Construct a frequency distribution from this histogram. Use a first class of 625–675.

19. *Response Time* Use the frequency polygon below to answer the following questions.



a) How many calls were responded to in 5 minutes?
b) How many calls were responded to in 6 minutes or less?
c) How many calls were included in the survey?
d) Construct a frequency distribution from the frequency polygon.
e) Construct a histogram from the frequency distribution in part (d).

20. *San Diego Zoo* Use the frequency polygon below to answer the following questions.



a) How many families visited the San Diego Zoo four times?
b) How many families visited the San Diego Zoo at least six times?
c) How many families were surveyed?
d) Construct a frequency distribution from the frequency polygon.
e) Construct a histogram from the frequency distribution in part (d).

21. Construct a histogram and a frequency polygon from the frequency distribution given in Exercise 7 of Section 13.3. See page 765.

22. Construct a histogram and a frequency polygon from the frequency distribution given in Exercise 8 of Section 13.3. See page 765.

23. ***College Credits*** Eighteen students in a geology class were asked how many college credits they had earned. The responses are as follows. Construct a stem-and-leaf display.

10	15	24	36	48	45
42	53	60	17	24	30
33	45	48	62	54	60

24. ***Distance to Work*** Twenty workers at a small company were asked how many miles they drive to work, one way. The responses are as follows. Construct a stem-and-leaf display. For single-digit data, use a stem of 0.

12	18	3	8	12	25	21
33	15	2	5	27	41	22
19	13	23	34	17	16	

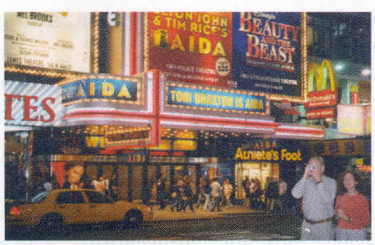
25. ***Starting Salaries*** Starting salaries (in thousands of dollars) for social workers with a bachelor of science degree and no experience are shown for a random sample of 25 different social workers.

27	28	29	31	33
28	28	29	31	33
28	28	30	32	33
28	29	30	32	34
28	29	30	32	34

a) Construct a frequency distribution. Let each class have a width of one.
b) Construct a histogram.
c) Construct a frequency polygon.
d) Construct a stem-and-leaf display.

26. ***Broadway Shows*** The ages of a random sample of people attending a Broadway show are

20	23	25	30	32	35	39	44
21	23	26	30	33	35	40	45
21	24	27	30	34	35	40	45
22	24	28	31	34	37	40	46
23	25	28	31	34	38	42	47



Broadway in New York City

a) Construct a frequency distribution with a first class of 20–24.
b) Construct a histogram.
c) Construct a frequency polygon.
d) Construct a stem-and-leaf display.

27. ***Advertising*** The following table shows the 50 leading companies in terms of dollars spent for advertising in 2001, in the United States, rounded to the nearest million dollars.

Company	Advertising Spending (millions)
General Motors Corp.	$3374
Procter & Gamble Co.	$2541
Ford Motor Co.	$2408
PepsiCo	$2210
Pfizer	$2190
DaimlerChrysler	$1985
AOL Time Warner	$1885
Philip Morris Cos.	$1816
Walt Disney Co.	$1757
Johnson & Johnson	$1618
Unilever	$1484
Sears, Roebuck & Co.	$1480
Verizon Communications	$1462
Toyota Motor Corp.	$1399
AT&T Corp.	$1372
Sony Corp.	$1310
Viacom	$1283
McDonald's Corp.	$1195
Diageo	$1181
Sprint Corp.	$1160
Merck & Co.	$1137
Honda Motor Co.	$1103
J.C. Penney Corp.	$1086
U.S. Government	$1057
L'Oreal	$1041
IBM Corp.	$994
Bristol-Myers Squibb Co.	$974
Nestlé	$967
SBC Communications	$943
Target Corp.	$926
Microsoft Corp.	$920
Coca-Cola Co.	$904
Hewlett-Packard Co.	$899
AT&T Wireless	$888
General Mills	$884
GlaxoSmithKline	$881
WorldCom	$840
Sara Lee Corp.	$812
Home Depot	$778
Nissan Motor Co.	$775

(continued on next page)

Company	Advertising Spending (millions)
Wyeth	$771
Estee Lauder Cos.	$766
Federated Department Stores	$746
Yum Brands	$677
News Corp.	$670
ConAgra	$668
General Electric Co.	$664
Anheuser-Busch Cos.	$656
Mars Inc.	$615
Kmart Corp.	$597

Source: AdAge.com

a) Construct a frequency distribution with the first class $597 million–$905 million.
b) Construct a histogram.
c) Construct a frequency polygon.

28. ***U.S. Ambassadors*** The ages of a random sample of U.S. ambassadors are

40	43	45	50	52	55	59	64
41	43	46	50	53	55	60	65
41	44	47	50	54	55	60	65
42	44	48	51	54	57	60	66
43	45	48	51	54	58	62	67

a) Construct a frequency distribution with the first class 40–44.
b) Construct a histogram.
c) Construct a frequency polygon.

Challenge Problems/Group Activities

29. ***Birthdays*** **a)** What do you believe a histogram of the months in which the students in your class were born (January is month 1 and December is month 12) would look like? Explain.
b) By asking, determine the month in which the students in your class were born (include yourself).
c) Construct a frequency distribution containing 12 classes.
d) Construct a histogram from the frequency distribution in part (c).
e) Construct a frequency polygon of the frequency distribution in part (c).

30. ***Social Security Numbers*** Repeat Exercise 29 for the last digit of the students' social security numbers. Include classes for the digits 0–9.

Internet/Research Activity

31. Over the years many changes have been made in the U.S. Social Security System.
a) Do research and determine the number of people receiving social security benefits for the years 1945, 1950, 1955, 1960, . . . , 2000. Then construct a frequency distribution and histogram of the data.
b) Determine the maximum amount that self-employed individuals had to pay into social security (the FICA tax) for the years 1945, 1950, 1955, 1960, . . . , 2000. Then construct a frequency distribution and a histogram of the data.

13.5 MEASURES OF CENTRAL TENDENCY

Most people have an intuitive idea of what is meant by an "average." The term is used daily in many familiar ways: "This car averages 19 miles per gallon." "The average test grade was 78." "The average height of adult males is 5 feet 9 inches."

An ***average*** is a number that is representative of a group of data. There are at least four different averages: the mean, the median, the mode, and the midrange. Each is calculated differently and may yield different results for the same set of data. Each will result in a number near the center of the data; for this reason, averages are commonly referred to as ***measures of central tendency***.

The ***arithmetic mean***, or simply the ***mean***, is symbolized either by $\bar{x}$ (read "x bar") or by the Greek letter mu, μ. The symbol $\bar{x}$ is used when the mean of a *sample* of the population is calculated. The symbol μ is used when the mean of the *entire population* is calculated. Unless otherwise indicated, we will assume that the data featured in this book represent samples; therefore, we will use $\bar{x}$ for the mean.

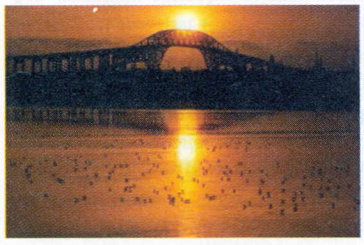
The Greek letter sigma, Σ, is used to indicate "summation." The notation Σx, read "the sum of x," is used to indicate the sum of all the data. For example, if there are five pieces of data, 4, 6, 1, 0, 5, then $\Sigma x = 4 + 6 + 1 + 0 + 5 = 16$.

DID YOU KNOW

Buying the American Dream

San Francisco, California

One of the biggest dreams for most people is to own their own home. Yet depending on where you live, the American dream may be hard to achieve. In 2002, San Francisco had the highest median home price for major metropolitan housing markets, $482,300. Beaumont/Port Arthur had the lowest median home price for major metropolitan housing markets, $76,800.



Beaumont/Port Arthur, Texas
Source: National Association of Realtors, Quarterly Housing Affordability Index

Now we can discuss the procedure for determining the mean of a set of data.

The mean, $\bar{x}$, is the sum of the data divided by the number of pieces of data. The formula for calculating the mean is

$$\bar{x} = \frac{\Sigma x}{n}$$

where Σx represents the sum of all the data and n represents the number of pieces of data.

The most common use of the word *average* is the mean.

EXAMPLE 1 *Find the Mean*

Find the mean age of a group of volunteers at an American Red Cross office if the ages of the individuals are 27, 18, 48, 34, and 48.

SOLUTION:

$$\bar{x} = \frac{\Sigma x}{n} = \frac{27 + 18 + 48 + 34 + 48}{5} = \frac{175}{5} = 35$$

Therefore, the mean, $\bar{x}$, is 35 years. ▲

The mean represents "the balancing point" of a set of data. For example, if a seesaw were pivoted at the mean and uniform weights were placed at points corresponding to the ages in Example 1, the seesaw would balance. Figure 13.13 shows the five ages given in Example 1 and the calculated mean.

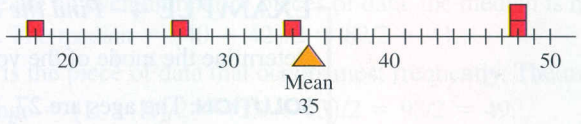


Figure 13.13

A second average is the *median.* To find the median of a set of data, *rank the data* from smallest to largest, or largest to smallest, and determine the value in the middle of the set of *ranked data.* This value will be the median.

The **median** is the value in the middle of a set of *ranked data.*

EXAMPLE 2 *Find the Median*

Determine the median of the volunteers' ages in Example 1.

SOLUTION: Ranking the data from smallest to largest gives 18, 27, 34, 48, and 48. Since 34 is the value in the middle of this set of ranked data (two pieces of data above it and two pieces below it) 34 years is the median. ▲

If there are an even number of pieces of data, the median will be halfway between the two middle pieces. In this case, to find the median, add the two middle pieces and divide this sum by 2.

EXAMPLE 3 *Find the Median of an Even Number of Pieces of Data*

Determine the median of the following sets of data.

a) 7, 12, 14, 15, 9, 14, 9, 10

b) 7, 8, 8, 8, 9, 10

SOLUTION:

a) Ranking the data gives 7, 9, 9, 10, 12, 14, 14, 15. There are eight pieces of data. Therefore, the median will lie halfway between the two middle pieces, the 10 and the 12. The median is $\frac{10 + 12}{2}$ or $\frac{22}{2}$ or 11.

b) There are six pieces of data and they are already ranked. Therefore, the median lies halfway between the two middle pieces. Both middle pieces are 8's. The median is $\frac{8 + 8}{2}$, or $\frac{16}{2}$, or 8. ▲

TIMELY TIP Data must be ranked before finding the median. A common error made when finding the median is neglecting to arrange the data in ascending (increasing) or in descending (decreasing) order.

A third average is the ***mode***.

The **mode** is the piece of data that occurs most frequently.

EXAMPLE 4 *Find the Mode*

Determine the mode of the volunteers' ages in Example 1.

SOLUTION: The ages are 27, 18, 48, 34, and 48. The age 48 is the mode because it occurs twice and the other values occur only once. ▲

If each piece of data occurs only once, the set of data has no mode. For example, the set of data 1, 2, 3, 4, 5 has no mode. If two values in a set of data occur more often than all the other data, we consider both these values as modes and say the data is **bimodal*** (which means two modes). For example, the set of data 1, 1, 2, 3, 3, 5 has two modes, 1 and 3.

The last average that we will discuss is the midrange. The ***midrange*** is the value halfway between the lowest (L) and highest (H) values in a set of data. It is found by adding the lowest and highest values and dividing the sum by 2. A formula for finding the midrange follows.

$$\textbf{Midrange} = \frac{\text{lowest value} + \text{highest value}}{2}$$

*Some textbooks say that sets of data such as 1, 1, 2, 3, 3, 5 have no mode.

DID YOU KNOW

When Babies' Eyes Are Smiling

Experimental psychologists formulate hypotheses about human behavior, design experiments to test them, make observations, and draw conclusions from their data. They use statistical concepts at each stage to help ensure that their conclusions are valid. In one experiment, researchers observed that 2-month-old infants who learned to move their heads so as to make a mobile turn began to smile as soon as the mobile turned. Babies in the control group did not smile as often when the mobile moved independently of their head turning. The researchers concluded that it was not the movement of the mobile that made the infant smile; rather, the infants smiled at their own achievement.

EXAMPLE 5 *Find the Midrange*

Determine the midrange of the volunteers' ages given in Example 1.

SOLUTION: The ages of the volunteers are 27, 18, 48, 34, and 48. The lowest age is 18 and the highest age is 48.

$$\text{Midrange} = \frac{\text{lowest} + \text{highest}}{2} = \frac{18 + 48}{2} = \frac{66}{2} = 33 \text{ years}$$

▲

The "average" of the ages 27, 18, 48, 34, 48 can be considered any one of the following values: 35 (mean), 34 (median), 48 (mode), or 33 (midrange). Which average do you feel is most representative of the ages? We will discuss this question later in this section.

EXAMPLE 6 *Measures of Central Tendency*

The salaries of eight selected teachers rounded to the nearest thousand dollars are 40, 25, 28, 35, 42, 60, 60, and 73. For this set of data, determine the (a) mean, (b) median, (c) mode, and (d) midrange, and then (e) rank the measures of central tendency from lowest to highest.

SOLUTION:

a) $\bar{x} = \dfrac{\Sigma x}{n} = \dfrac{40 + 25 + 28 + 35 + 42 + 60 + 60 + 73}{8} = \dfrac{363}{8} = 45.375$

b) Ranking the data from the smallest to largest gives

$$25, 28, 35, 40, 42, 60, 60, 73$$

Since there are an even number of pieces of data, the median is halfway between 40 and 42. The median $= (40 + 42)/2 = 82/2 = 41$.

c) The mode is the piece of data that occurs most frequently. The mode is 60.

d) The midrange $= (L + H)/2 = (25 + 73)/2 = 98/2 = 49$.

e) The averages from lowest to highest are the median, mean, midrange, and mode. Their values are 41, 45.375, 49, and 60, respectively. ▲

At this point you should be able to calculate the four measures of central tendency: mean, median, mode, and midrange. Now let's examine the circumstances in which each is used.

The mean is used when each piece of data is to be considered and "weighed" equally. It is the most commonly used average. It is the only average that can be affected by *any* change in the set of data; for this reason, it is the most sensitive of all the measures of central tendency (see Exercise 23).

Occasionally, one or more pieces of data may be much greater or much smaller than the rest of the data. When this situation occurs, these "extreme" values have the effect of increasing or decreasing the mean significantly so that the mean will not be representative of the set of data. Under these circumstances, the median should be used instead of the mean. The median is often used in describing average family incomes because a relatively small number of families have extremely large incomes. These few incomes would inflate the mean income, making it nonrepresentative of the millions of families in the population.

Consider a set of exam scores from a mathematics class: 0, 16, 19, 65, 65, 65, 68, 69, 70, 72, 73, 73, 75, 78, 80, 85, 88, 92. Which average would best represent these

grades? The mean is 64.06. The median is 71. Since only 3 of the 18 scores fall below the mean, the mean would not be considered a good representative score. The median of 71 probably would be the better average to use.

The mode is the piece of data, if any, that occurs most frequently. Builders planning houses are interested in the most common family size. Retailers ordering shirts are interested in the most common shirt size. An individual purchasing a thermometer might choose one, from those on display, whose temperature reading is the most common reading among those on display. These examples illustrate how the mode may be used.

The midrange is sometimes used as the average when the item being studied is constantly fluctuating. Average daily temperature, used to compare temperatures in different areas, is calculated by adding the lowest and highest temperatures for the day and dividing the sum by 2. The midrange is actually the mean of the high value and the low value of a set of data. Occasionally, the midrange is used to estimate the mean, since it is much easier to calculate.

Sometimes an average itself is of little value, and care must be taken in interpreting its meaning. For example, Jim is told that the average depth of Willow Pond is only 3 feet. He is not a good swimmer but decides it is safe to go out a short distance in this shallow pond. After he is rescued, he exclaims, "I thought this pond was only 3 feet deep." Jim didn't realize an average does not indicate extreme values or the spread of the values. The spread of data is discussed in Section 13.6.

Measures of Position

Measures of position are used to describe the position of a piece of data in relation to the rest of the data. If you took the Scholastic Aptitude Test (SAT) before applying to college, your score was described as a measure of position rather than a measure of central tendency. ***Measures of position*** are often used to make comparisons, such as comparing the scores of individuals from different populations, and are generally used when the amount of data is large.

Two measures of position are ***percentiles*** and ***quartiles***. There are 99 percentiles dividing a set of data into 100 equal parts; see Fig. 13.14. For example, suppose that you scored 490 on the math half of the SAT, and the score of 490 was reported to be in the 78th percentile of high school students. This wording *does not* mean that 78% of your answers were correct; it *does* mean that you outperformed about 78% of all those taking the exam. In general, a score in the nth percentile means that you outperformed about $n\%$ of the population who took the test and that $(100 - n)\%$ of the people taking the test performed better than you did.

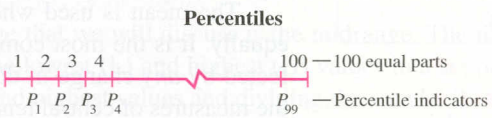


Figure 13.14

EXAMPLE 7 *English Achievement Test*

Kara Hopkins took an English achievement test to obtain college credit by exam for freshman English. Her score was at the 81st percentile. Explain what that means.

SOLUTION: If a score is at the 81st percentile, it means that about 81% of the scores are below that score. Therefore, Kara scored better than about 81% of the students taking the exam. Also, about 19% of all students taking the exam scored higher than she did. ▲

Quartiles are another measure of position. Quartiles divide data into four equal parts: The first quartile is the value that is higher than about 1/4 or 25% of the population. It is the same as the 25th percentile. The second quartile is the value that is higher than about 1/2 the population and is the same as the 50th percentile, or the median. The third quartile is the value that is higher than about 3/4 of the population and is the same as the 75th percentile; see Fig. 13.15.

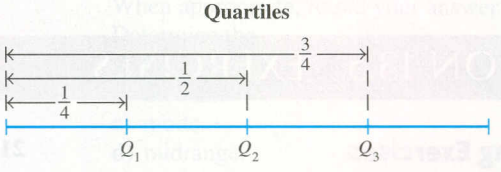


Figure 13.15

To Find the Quartiles of a Set of Data:

1. Order the data from smallest to largest.
2. Find the median, or the 2nd quartile, of the set of data. If there are an odd number of pieces of data, the median is the middle value. If there are an even number of pieces of data, the median will be halfway between the two middle pieces of data.
3. The first quartile, Q_1, is the median of the lower half of the data; that is, Q_1 is the median of the data less than Q_2.
4. The third quartile, Q_3, is the median of the upper half of the data; that is, Q_2 is the median of the data greater than Q_2.

EXAMPLE 8 *Finding Quartiles*

Electronics World is concerned about the high turnover of its sales staff. A survey was done to determine how long (in months) their sales staff had been in their current positions. The responses of 27 sales staff follow. Determine Q_1, Q_2, and Q_3.

25	3	7	15	31	36	17	21	2
11	42	16	23	19	21	9	20	5
8	12	27	14	39	24	18	6	10

SOLUTION: First we order the data from smallest to largest.

2	3	5	6	7	8	9	10	11
12	14	15	16	17	18	19	20	21
21	23	24	25	27	31	36	39	42

Next we find the median. Since there are 27 pieces of data, an odd number, the median will be the middle value. The middle value is 17, with 13 pieces of data less than 17 and 13 pieces of data greater than 17. Therefore, the median, Q_2, is 17, shown in red.

To find Q_1, the median of the lower half of the data, we need to find the median of the 13 pieces of data that are less than Q_2. The middle value of the lower half of the data is 9. There are 6 pieces of data less than 9 and 6 pieces of data greater than 9. Therefore, Q_1 is 9, shown in blue.

To find Q_3, the median of the upper half of the data, we need to find the median of the 13 pieces of data that are greater than 17, or Q_2. The middle value of the upper half of the data is 24. There are 6 pieces of data greater than 17 but less than 24 and 6 pieces of data greater than 24. Therefore, Q_3 is 24, shown in blue. ▲

SECTION 13.5 EXERCISES

Concept/Writing Exercises

1. What is a set of *ranked data*?
2. Describe the mean of a set of data and explain how to find it.
3. Describe the median of a set of data and explain how to find it.
4. Describe the midrange of a set of data and explain how to find it.
5. Describe the mode of a set of data and explain how to find it.
6. When might the mode be the preferred average to use? Give an example.
7. When might the median be the preferred average to use? Give an example.
8. When might the midrange be the preferred average to use? Give an example.
9. When might the mean be the preferred average to use? Give an example.
10. **a)** What symbol is used for the sample mean?
    b) What symbol is used for the population mean?

Practice the Skills

In Exercises 11–20, determine the mean, median, mode, and midrange of the set of data. Round your answer to the nearest tenth.

11. 5, 6, 6, 9, 10, 10, 10, 20, 23

12. 9, 10, 15, 17, 15, 14, 370, 45, 42, 13

13. 66, 72, 84, 45, 90, 42, 86

14. 7, 5, 8, 8, 8, 10, 12

15. 1, 3, 5, 7, 9, 11, 13, 15

16. 40, 50, 30, 60, 90, 100, 140

17. 1, 7, 11, 27, 36, 14, 12, 9, 1

18. 1, 1, 1, 1, 4, 4, 4, 4, 6, 8, 10, 12, 15, 21

19. 6, 8, 12, 13, 11, 13, 15, 17

20. 5, 15, 5, 15, 5, 15

21. ***Best-Seller List*** The number of weeks the top 10 hardcover fiction novels were on the best-seller list as of July 28, 2002, is 2, 3, 3, 5, 19, 7, 4, 5, 11, 6. Find the mean, median, mode, and midrange.

22. ***Daily Tips*** The amount of money Amy Striegel collected in tips as a waitress in each of seven days is $25, $60, $37, $48, $100, $140, $59. Find the mean, median, mode, and midrange.

23. ***Change in the Data*** The mean is the "most sensitive" average because it is affected by any change in the data.
a) Determine the mean, median, mode, and midrange for 1, 2, 3, 5, 5, 7, 11.
b) Change the 7 to a 10 in part (a). Find the mean, median, mode, and midrange.
c) Which averages were affected by changing the 7 to a 10?
d) Which averages will be affected by changing the 11 to a 10 in part (a)?

24. ***Life Expectancy*** In 2000, the National Center for Health Statistics indicated a new record "average life expectancy" of 76.9 years for the total U.S. population. The average life expectancy for men was 74.1 years, and for women it was 79.5 years. Which "average" do you think the National Center for Health is using? Explain your answer.

25. ***A Grade of B*** To get a grade of B, a student must have a mean average of 80 or greater. Jim Condor has a mean average of 79 for 10 quizzes. He approaches his teacher and

asks for a B, reasoning that he missed a B by only one point. What is wrong with Jim's reasoning?

26. ***Employee Salaries*** The salaries of 10 employees of a small company follow.

$28,000	$64,000
25,000	24,000
31,000	27,000
26,000	81,000
26,000	29,000

Calculate the

a) mean.
b) median.
c) mode.
d) midrange.
e) If the employees wanted to demonstrate the need for a raise, which average would they use to show they are being underpaid: the mean or the median? Explain.
f) If the management did not want to give the employees a raise, which average would they use: the mean or the median? Explain.

27. ***National Parks*** The 10 national parks and recreation areas with the most visitors in 2001 are listed below.

Park	Number of Visitors (millions of people)
Blue Ridge Parkway	19.7
Golden Gate National Recreation Area	13.4
Great Smoky Mountains National Park	9.5
Lake Mead National Recreation Area	8.8
Gateway National Recreation Area	8.2
George Washington Memorial Parkway	7.8
Natchez Trace Parkway	5.5
Statue of Liberty National Monument	5.4
Delaware Water Gap National Recreation Area	4.8
Castle Clinton	4.6

Source: National Park Service (www.nps.gov.)

Rounding your answers to the nearest tenth, determine the

a) mean.
b) median.
c) mode.
d) midrange.

28. ***Living Expenses*** Bob Exler's monthly living expenses for 1 year are as follows:

$1000	$850	$1370	$1400
1900	850	1350	1250
1600	900	1110	1230

When appropriate, round your answer to the nearest cent. Determine the

a) mean.
b) median.
c) mode.
d) midrange.

29. ***Costliest Hurricanes*** The damage caused, in billions of dollars, by the 11 costliest hurricanes in the United States is listed below.

Hurricane	Year	Cost (billions)
Andrew	1992	$26.5
Hugo	1989	7.0
Floyd	1999	4.5
Fran	1996	3.2
Opal	1995	3.0
Georges	1998	2.3
Frederic	1979	2.3
Agnes	1972	2.1
Alicia	1983	2.0
Bob	1991	1.5
Juan	1985	1.5

Source: National Oceanic and Atmospheric Administration

Rounding your answers to the nearest tenth, determine the

a) mean.
b) median.
c) mode.
d) midrange.
e) Which average is the best measure of central tendency for this set of data? Explain.

30. ***Exam Average*** Malcolm Sander's mean average on five exams is 76. Find the sum of his scores.

31. ***Exam Average*** Jeremy Urban's mean average on six exams is 85. Find the sum of his scores.

32. ***Creating a Data Set*** Construct a set of five pieces of data in which the mode has a lower value than the median and the median has a lower value than the mean.

33. ***Creating a Data Set*** Construct a set of six pieces of data with a mean, median, and midrange of 75 and where no two pieces of data are the same.

34. ***Creating a Data Set*** Construct a set of six pieces of data with a mean of 88 and where no two pieces of data are the same.

35. ***Water Park*** For the 2003 season, 24,000 people visited the Blue Lagoon Water Park. The park was open 120 days for water activities. The highest number of visitors on a single day was 500. The lowest number of visitors on a single day was 50. Determine whether it is possible to find the following with the given information:
a) the mean number of visitors per day.
b) the median number of visitors per day.
c) the mode number of visitors per day.
d) the midrange number of visitors per day.
e) Find all the measures of central tendency that can be found with the information and explain why the others cannot be found.

36. ***Determine a Necessary Grade*** A mean average of 80 or greater for five exams is needed for a final grade of B. Jorge Rivera's first four exam grades are 73, 69, 85, and 80. What grade does Jorge need on the fifth exam to get a B in the course?

37. ***Grading Methods*** A mean average of 60 on seven exams is needed to pass a course. On her first six exams, Sheryl Ward received grades of 49, 72, 80, 60, 57, and 69.
a) What grade must she receive on her last exam to pass the course?
b) An average of 70 is needed to get a C in the course. Is it possible for Sheryl to get a C? If so, what grade must she receive on the seventh exam?
c) If her lowest grade of the exams already taken is to be dropped, what grade must she receive on her last exam to pass the course?
d) If her lowest grade of the exams already taken is to be dropped, what grade must she receive on her last exam to get a C in the course?

38. ***Central Tendencies*** Which of the measures of central tendency *must* be an actual piece of data in the distribution? Explain.

39. ***Creating a Data Set*** Construct a set of six pieces of data such that if only one piece of data is changed, the mean, median, and mode will all change.

40. ***Changing One Piece of Data*** Consider the set of data 1, 1, 1, 2, 2, 2. If one 2 is changed to a 3, which of the following will change: mean, median, mode, midrange? Explain.

41. ***Changing One Piece of Data*** Is it possible to construct a set of six different pieces of data such that by changing only one piece of data you cause the mean, median, mode, and midrange to change? Explain.

42. ***Grocery Expenses*** The Taylor's have recorded their weekly grocery expenses for the past 12 weeks and determined the mean weekly expense was \$85.20. Later Mrs. Taylor discovered 1 week's expense of \$74 was incorrectly recorded as \$47. What is the correct mean?

43. ***Percentiles*** For any set of data, what must be done to the data before percentiles can be determined?

44. ***Percentiles*** Josie Waverly scored in the 73rd percentile on the verbal part of her College Board test. What does that mean?

45. ***Percentiles*** When a national sample of heights of kindergarten children was taken, Kevin was told that he was in the 35th percentile. Explain what that means.

46. ***Percentiles*** A union leader is told that, when all workers' salaries are considered, the first quartile is \$20,750. Explain what that means.

47. ***Quartiles*** The prices of the 21 top-rated 27-inch direct-view television sets, as rated in the March 2002 issue of *Consumer Reports*, are as follows:

\$290	\$330	\$350	\$400	\$450	\$650	\$700
300	350	350	430	500	650	750
300	350	350	450	600	700	800

Determine
a) Q_2. **b)** Q_1. **c)** Q_3.

48. ***Quartiles*** The cost, in cents, per half-cup serving of the top 20 rated brands of vanilla ice cream, as reported in the July 2002 issue of *Consumer Reports*, are as follows:

17	21	27	27	28	33	80
17	24	27	28	28	38	81
20	25	27	28	31	74	

Determine
a) Q_2. **b)** Q_1. **c)** Q_3.

49. ***The 50th Percentile*** Give the names of two other statistics that have the same value as the 50th percentile.

50. ***College Admissions*** Jonathan took an admission test for the University of California and scored in the 85th percentile. The following year Jonathan's sister Kendra took a similar admission test for the University of California and scored in the 90th percentile.
a) Is it possible to determine which of the two answered the higher percent of questions correctly on their respective exams? Explain your answer.

b) Is it possible to determine which of the two was in a better relative position with regard to their respective populations? Explain.

51. ***Employee Salaries*** The following statistics represent weekly salaries at the Midtown Construction Company:

Mean	$510	First quartile	$470
Median	$500	Third quartile	$535
Mode	$490	83rd percentile	$575

a) What is the most common salary?
b) What salary did half the employees' salaries surpass?
c) About what percent of employees' salaries surpassed $535?
d) About what percent of employees' salaries were less than $470?
e) About what percent of employees' salaries surpassed $575?
f) If the company has 100 employees, what is the total weekly salary of all employees?

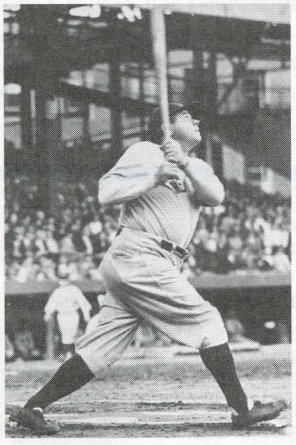
Challenge Problems/Group Activities

52. ***The Mean of the Means*** Consider the following five sets of values.

i) 5 6 7 7 8 9 14
ii) 3 6 8 9
iii) 1 1 1 2 5
iv) 6 8 9 12 15
v) 50 51 55 60 80 100

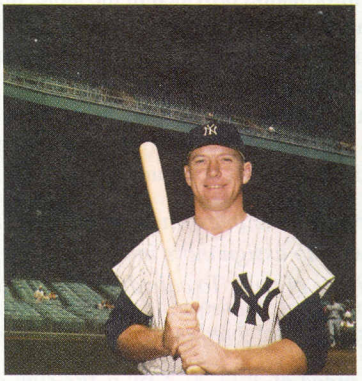
a) Compute the mean of each of the five sets of data.
b) Compute the mean of the five means in part (a).
c) Find the mean of the 27 pieces of data.
d) Compare your answer in part (b) to your answer in part (c). Are the values the same? Does your answer make sense? Explain.

53. ***Ruth versus Mantle*** The tables to the right compare the batting performances for selected years for two well-known former baseball players, Babe Ruth and Mickey Mantle.



Babe Ruth
Boston Red Sox 1914–1919
New York Yankees 1920–1934

Year	At Bats	Hits	Pct.
1925	359	104	
1930	518	186	
1933	459	138	
1916	136	37	
1922	406	128	
Total	1878	593	



Mickey Mantle
New York Yankees 1951–1968

Year	At Bats	Hits	Pct.
1954	543	163	
1957	474	173	
1958	519	158	
1960	527	145	
1962	377	121	
Total	2440	760	

a) For each player, compute the batting average percent (pct.) for each year by dividing the number of hits by the number of at bats. Round off to the nearest thousandth. Place the answers in the pct. column.
b) Going across each of the five horizontal lines (for example Ruth, 1925, vs. Mantle, 1954), compare the percents (pct.) and determine which is greater in each case.
c) For each player, compute the mean batting average percent for the 5 given years by dividing the total hits by the total at bats. Which is greater, Ruth's or Mantle's?
d) Based on your answer in part (b), does your answer in part (c) make sense? Explain.
e) Find the mean percent for each player by adding the five pcts. and dividing by 5.Which is greater, Ruth's or Mantle's?
f) Why do the answers obtained in parts (c) and (e) differ? Explain.
g) Who would you say has the better batting average percent for the 5 years selected? Explain.

54. ***Employee Salaries*** The following table gives the annual salary distribution for employees at Kulzer's Home Improvement.

Annual Salary	Number Receiving Salary
$100,000	1
85,000	2
24,000	6
21,000	4
18,000	5
17,000	7

Using the information provided in the table, find the

a) mean annual salary.
b) median annual salary.
c) mode annual salary.
d) midrange annual salary.
e) Which is the best measure of central tendency for this set of data? Explain your answer.

Weighted Average *Sometimes when we wish to find an average, we may wish to assign more importance, or weight, to some of the pieces of data. To calculate a* ***weighted average,*** *we use the formula:* $\text{weighted average} = \dfrac{\Sigma xw}{\Sigma w}$, *where* w *is the weight of the piece of data,* x; Σxw *is the sum of the products of each piece of data multiplied by its weight; and* Σw *is the sum of the weights. For example, suppose that students in a class need to submit a report that counts 20% of their grade, they need to take a midterm exam that counts 30% of their grade, and they need to take a final exam that counts 50% of their grade. Suppose that a student got a 72 on the report, an 85 on the midterm exam, and a 93 on the final exam. To determine this student's weighted average, first find* Σxw: $\Sigma xw = 72(0.20) + 85(0.30) + 93(0.50) = 86.4$. *Next find* Σw, *the sum of the weights:* $\Sigma w = 0.20 + 0.30 + 0.50 = 1.00$. *Now determine the weighted average as follows*

$$\text{weighted average} = \frac{\Sigma xw}{\Sigma w} = \frac{86.4}{1.00} = 86.4$$

Thus, the weighted average is 86.4. Note that Σw *does not always have to be 1.00. In Exercise 55 and 56, use the weighted average formula.*

55. ***Course Average*** Suppose that your final grade for a course is determined by a midterm exam and a final exam. The midterm exam is worth 40% of your grade and the final exam is worth 60%. If your midterm exam grade is 84 and your final exam grade is 94, calculate your final average.

56. ***Grade Point Average*** In a four-point grade system, an A corresponds to 4.0 points, a B corresponds to a 3.0 points, a C corresponds to a 2.0 points, and a D corresponds to a 1.0 points. No points are awarded for an F. Last semester Tanya Reeves received a B in a four-credit hour course, an A in a three-credit hour course, a C in a three-credit hour course, and an A in another three-credit hour course. Grade point average (GPA) is calculated as a weighted average using the credit hours as weights and the number of points corresponding to the grade as pieces of data. Calculate Tanya's GPA for the previous semester. (Round your answer to the nearest hundredth.)

Recreational Mathematics

57. ***Your Exam Average*** **a)** Calculate the mean, median, mode, and midrange of your exam grades in your mathematics course.
b) Which measure of central tendency best represents your average grade?
c) Which measure of central tendency would you rather use as your average grade?

58. ***Purchases*** Matthew Abbott purchased some items at Staples each day for five days. The mode of the number of items Matthew purchased is higher than the median of the number of items he purchased. The median of the number of items Matthew purchased is higher than the mean of the number of items he purchased. Each day he purchased at least two items but no more than seven items.
a) How many items did Matthew purchase each day? (*Note:* There is more than one correct answer.)
b) Determine the mean, median, and mode for your answer to part (a).

Internet/Research Activity

59. Two other measures of location that we did not mention in this section are *stanines* and *deciles.* Use statistics books, books on educational testing and measurements, and Internet websites to write a report on what stanines and deciles are and when percentiles, quartiles, stanines, and deciles are used.

13.6 MEASURES OF DISPERSION

The measures of central tendency by themselves do not always give sufficient information to analyze a situation and make decisions. As an example, two manufacturers of airplane engines are being considered for a contract. Manufacturer A's engines have an average (mean) life of 1000 hours of flying time before they must be rebuilt. Manufacturer B's engines have an average life of 950 hours of flying time before they

must be rebuilt. If you assume that both cost the same, which engines should be purchased? The average engine life may not be the most important factor. The fact that manufacturer A's engines have an average life of 1000 hours could mean that half will last about 500 hours and the other half will last about 1500 hours. If in fact all manufacturer B's engines have a life span of between 900 and 1000 hours, then all of manufacturer B's engines are more consistent and reliable. If A's engines were purchased, they would all have to be rebuilt every 300 hours or so because it would be impossible to determine which ones would fail first. If B's engines were purchased, they could go much longer before having to be rebuilt. This example is of course an exaggeration used to illustrate the importance of knowing something about the *spread,* or *variability,* of the data.

Measures of dispersion are used to indicate the spread of the data. The range and standard deviation* are the measures of dispersion that will be discussed in this book.

The ***range*** is the difference between the highest and lowest values; it indicates the total spread of the data.

Range = highest value − lowest value

EXAMPLE 1 *Find the Range*

Twelve different foods were selected and the amount of potassium, in milligrams, in each was recorded. Find the range of the following amounts of potassium.

900, 789, 400, 408, 860, 780, 451, 502, 496, 503, 555, 566

SOLUTION: Range = highest value − lowest value = 900 − 400 = 500. The range of the amounts of potassium is 500 mg. ▲

The second measure of dispersion, the ***standard deviation***, measures how much the data *differ from the mean.* It is symbolized either by the letter s or by the Greek letter sigma, σ.† The s is used when the standard deviation of a *sample* is calculated. The σ is used when the standard deviation of the entire *population* is calculated. Since we are assuming that all data presented in this section are for samples, we use s to represent the standard deviation (note, however, that on the doctors' charts on page 794, σ is used. Also, we will use σ in the next section when we find standard scores.) The larger the spread of the data about the mean, the larger is the standard deviation. Consider the following two sets of data.

5, 8, 9, 10, 12, 13 8, 9, 9, 10, 10, 11

Both have a mean of 9.5. Which set of values on the whole do you believe differs less from the mean of 9.5? Figure 13.16 may make the answer more apparent. The scores in the second set of data are closer to the mean and therefore have a smaller standard deviation. You will soon be able to verify such relationships yourself.

Sometimes only a very small standard deviation is desirable or acceptable. Consider a cereal box that is to contain 8 oz of cereal. If the amount of cereal put into the

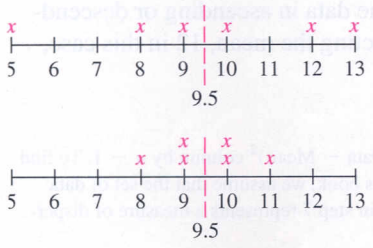


Figure 13.16

*Variance, another measure of dispersion, is the square of the standard deviation.

†Our alphabet uses both uppercase and lowercase letters, for example, A and a. The Greek alphabet also uses both uppercase and lowercase letters. The symbol Σ is the capital Greek letter sigma, and σ is the lowercase Greek letter sigma.

boxes varies too much—sometimes underfilling, sometimes overfilling—the manufacturer will soon be in trouble with consumer groups and government agencies.

At other times, a larger spread of data is desirable or expected. For example, intelligence quotients (IQs) are expected to exhibit a considerable spread about the mean, since everyone is different. The following procedure explains how we determine the standard deviation of a set of data.

To Find the Standard Deviation of a Set of Data:

1. Find the mean of the set of data.
2. Make a chart having three columns:

Data	Data − Mean	$(\text{Data} - \text{Mean})^2$

3. List the data vertically under the column marked Data.
4. Subtract the mean from each piece of data and place the difference in the Data − Mean column.
5. Square the values obtained in the Data − Mean column and record these values in the $(\text{Data} - \text{Mean})^2$ column.
6. Determine the sum of the values in the $(\text{Data} - \text{Mean})^2$ column.
7. Divide the sum obtained in step 6 by $n - 1$, where n is the number of pieces of data.*
8. Determine the square root of the number obtained in step 7. This number is the standard deviation of the set of data.

Example 2 illustrates the procedure to follow to find the standard deviation of a set of data.

EXAMPLE 2 *Find the Standard Deviation*

A veterinarian in an animal hospital recorded the following life spans of selected Labrador retrievers (to the nearest year):

$$7, 9, 11, 15, 18, 12$$

Find the standard deviation of the ages.

SOLUTION: First determine the mean:

$$\bar{x} = \frac{\Sigma x}{n} = \frac{7 + 9 + 11 + 15 + 18 + 12}{6} = \frac{72}{6} = 12$$

Next construct a table with three columns, as illustrated in Table 13.4, and list the data in the first column (it is often helpful to list the data in ascending or descending order). Complete the second column by subtracting the mean, 12 in this case, from each piece of data in the first column.

*To find the standard deviation of a sample divide the sum of $(\text{Data} - \text{Mean})^2$ column by $n - 1$. To find the standard deviation of a population divide the sum by n. In this book, we assume that the set of data represents a sample and divide by $n - 1$. The quotient obtained in step 7 represents a measure of dispersion called the *variance*.

DID YOU KNOW

Statistics and Opera Houses

Sydney Opera House in Sydney, Austrailia

Architects have developed a mathematical rule based on statistics to help them construct opera houses with exceptional acoustics. The rule was first developed by having conductors rate the overall sound quality in 23 opera houses. Then acoustical engineers measured several acoustical properties in those 23 buildings. By using statistical analysis, the engineers were able to determine which combination of properties produced exceptional sound and which of the acoustic characteristics were most important. This mathematical rule is now used in the development of new opera houses.

TABLE 13.4

Data	Data − Mean	(Data − Mean)^2
7	$7 - 12 = -5$	
9	$9 - 12 = -3$	
11	$11 - 12 = -1$	
12	$12 - 12 = 0$	
15	$15 - 12 = 3$	
18	$18 - 12 = 6$	
	0	

The sum of the values in the Data − Mean column should always be zero; if not, you have made an error. (If the mean is a decimal number, there may be a slight rounding error.)

Next square the values in the second column and place the squares in the third column (Table 13.5).

TABLE 13.5

Data	Data − Mean	(Data − Mean)^2
7	−5	$(-5)^2 = (-5)(-5) = 25$
9	−3	$(-3)^2 = (-3)(-3) = 9$
11	−1	$(-1)^2 = (-1)(-1) = 1$
12	0	$(0)^2 = (0)(0) = 0$
15	3	$(3)^2 = (3)(3) = 9$
18	6	$(6)^2 = (6)(6) = 36$
	0	80

Add the squares in the third column. In this case, the sum is 80. Divide this sum by one less than the number of pieces of data $(n - 1)$. In this case, the number of pieces of data is 6. Therefore, we divide by 5 and get

$$\frac{80}{5} = 16^*$$

Finally, take the square root of this number. Since $\sqrt{16} = 4$, the standard deviation, symbolized s, is 4. ▲

Now we will develop a formula for finding the standard deviation of a set of data. If we call the individual data x and the mean $\bar{x}$, we could write the three column heads Data, Data − Mean, and (Data − Mean)^2 in Table 13.4 as

x	$x - \bar{x}$	$(x - \bar{x})^2$

Let's follow the procedure we used to obtain the standard deviation in Example 2. We found the sum of the (Data − Mean)^2 column, which is the same as the sum of the $(x - \bar{x})^2$ column. We can represent the sum of the $(x - \bar{x})^2$ column by using the summation notation, $\Sigma(x - \bar{x})^2$. Thus, in Table 13.5, $\Sigma(x - \bar{x})^2 = 80$. We then divided this number by 1 less than the number of pieces of data, $n - 1$. Thus, we have

$$\frac{\Sigma(x - \bar{x})^2}{n - 1}$$

*16 is the variance, symbolized s^2, of this set of data.

Finally, we took the square root of this value to obtain the standard deviation.

Standard Deviation

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

EXAMPLE 3 *Find the Standard Deviation of Stock Prices*

The following are the prices of nine stocks on the New York Stock Exchange. Find the standard deviation of the prices.

$15, $28, $32, $36, $50, $52, $68, $74, $104

SOLUTION: The mean, $\bar{x}$, is

$$\bar{x} = \frac{\Sigma x}{n} = \frac{15 + 28 + 32 + 36 + 50 + 52 + 68 + 74 + 104}{9} = \frac{459}{9} = 51$$

The mean is $51.

TABLE 13.6

x	$x - \bar{x}$	$(x - \bar{x})^2$
15	−36	1296
28	−23	529
32	−19	361
36	−15	225
50	−1	1
52	1	1
68	17	289
74	23	529
104	53	2809
	0	6040

Table 13.6 shows us that $\Sigma(x - \bar{x})^2 = 6040$. Since there are nine pieces of data, $n - 1 = 9 - 1$, or 8.

$$s = \sqrt{\frac{\Sigma(x - \bar{x})^2}{n - 1}} = \sqrt{\frac{6040}{8}} = \sqrt{755} \approx 27.5$$

The standard deviation, to the nearest tenth, is $27.5. ▲

Standard deviation will be used in Section 13.7 to find the percent of data between any two values in a normal curve. Standard deviations are also often used in determining norms for a population (see Exercise 31).

SECTION 13.6 EXERCISES

Concept/Writing Exercises

1. Explain how to find the range of a set of data.
2. What does the standard deviation of a set of data measure?
3. Explain how to find the standard deviation of a set of data.
4. What is the standard deviation of a set of data in which all the data values are the same? Explain.
5. Why is measuring dispersion in observed data important?
6. What symbol is used to represent the sample standard deviation?
7. What symbol is used to represent the population standard deviation?
8. Can you think of any situations in which a large standard deviation may be desirable? Explain.
9. Can you think of any situations in which a small standard deviation may be desirable? Explain.
10. Without actually doing the calculations, decide which of the following two sets of data will have the greater standard deviation. Explain why.

    13, 16, 17, 18, 20, 24 16, 17, 17, 18, 18, 19

11. Without actually doing the calculations, decide which, if either, of the following two sets of data will have the greater standard deviation. Explain why.

    2, 4, 6, 8, 10 102, 104, 106, 108, 110

12. By studying the standard deviation formula, explain why the standard deviation of a set of data will always be greater than or equal to 0.
13. Patricia Wolff has two statistics classes, one in the morning and the other in the evening. On the midterm exam, the morning class had a mean of 75.2 and a standard deviation of 5.7. The evening class had a mean of 75.2 and a standard deviation of 12.5.
    a) How do the means compare?
    b) If we compare the set of scores from the first class with those in the second class, how will the distributions of the two sets of scores compare? Explain.
14. Explain why the standard deviation is usually a better measure of dispersion than the range.

Practice the Skills

In Exercises 15–22, determine the range and standard deviation of the set of data. When appropriate, round standard deviations to the nearest hundredths.

15. 7, 5, 2, 8, 13

16. 10, 10, 14, 16, 8, 8

17. 120, 121, 122, 123, 124, 125, 126

18. 3, 7, 8, 12, 0, 9, 11, 12, 6, 2

19. 4, 8, 9, 11, 13, 15

20. 9, 9, 9, 9, 9, 9, 9

21. 7, 9, 7, 9, 9, 10, 12

22. 52, 50, 54, 59, 40, 43, 64, 62

23. *Computer Games* Find the range and standard deviation of the following prices of selected computer games: $28, $28, $50, $45, $30, $45, $48, $18, $45, $23.

24. *Years until Retirement* Seven employees at a large company were asked the number of additional years they planned to work before retirement. Their responses were 10, 23, 28, 4, 1, 6, 12. Find the range and standard deviation of the number of years.

25. *Fishing Poles* Find the range and standard deviation of the following prices of selected fishing poles: $50, $120, $130, $60, $55, $75, $200, $110, $125, $175.

26. *Holiday Gifts* The amount of money seven college students planned to spend on gifts during the holiday season are as follows: $60, $100, $85, $35, $250, $150, $300. Find the range and standard deviation of the amounts.

Problem Solving

27. *Count Your Money* Six people were asked to determine the amount of money they were carrying, to the nearest dollar. The results were

$32, $60, $14, $25, $5, $68

a) Determine the range and standard deviation of the amounts.
b) Add $10 to each of the six amounts. How do you expect the range and standard deviation of the new set of data to change? Explain your answer.
c) Determine the range and standard deviation of the new set of data. Do the results agree with your answer to part (b)? If not, explain why.

28. a) *Adding to or Subtracting from Each Number* Pick any five numbers. Compute the mean and the standard deviation of this distribution.

b) Add 20 to each of the numbers in your original distribution and compute the mean and the standard deviation of this new distribution.

c) Subtract 5 from each number in your original distribution and compute the mean and standard deviation of this new distribution.

d) What conclusions can you draw about changes in the mean and the standard deviation when the same number is added to or subtracted from each piece of data in a distribution?

e) How will the mean and standard deviation of the numbers 6, 7, 8, 9, 10, 11, 12 differ from the mean and standard deviation of the numbers 596, 597, 598, 599, 600, 601, 602? Find the mean and standard deviation of both sets of numbers.

29. a) ***Multiplying Each Number*** Pick any five numbers. Compute the mean and standard deviation of this distribution.

b) Multiply each number in your distribution by 3 and compute the mean and the standard deviation of this new distribution.

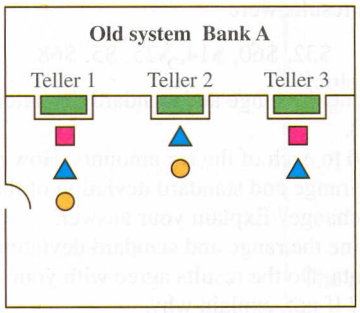
c) Multiply each number in your original distribution by 9 and compute the mean and the standard deviation of this new distribution.

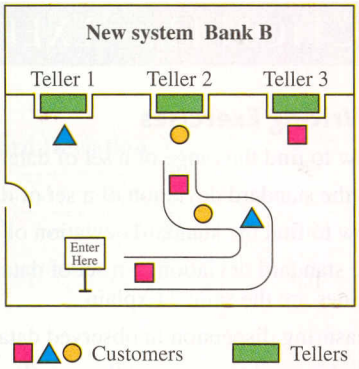
d) What conclusions can you draw about changes in the mean and the standard deviation when each value in a distribution is multiplied by the same number?

e) The mean and standard deviation of the distribution 1, 3, 4, 4, 5, 7 are 4 and 2, respectively. Use the conclusion drawn in part (d) to determine the mean and standard deviation of the distribution

5, 15, 20, 20, 25, 35

30. ***Waiting in Line*** Consider the following illustrations of two bank-customer waiting systems.

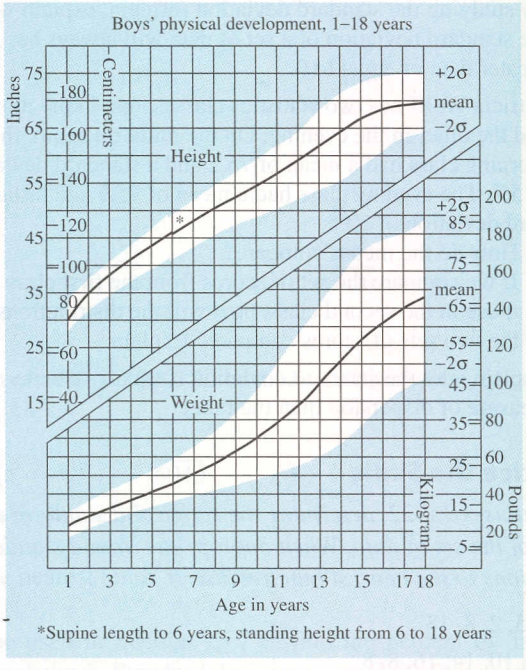




a) How would you expect the mean waiting time in Bank A to compare with the mean waiting time in Bank B? Explain your answer.

b) How would you expect the standard deviation of waiting times in Bank A to compare with the standard deviation of waiting times in Bank B? Explain your answer.

31. ***Height and Weight Distribution*** The chart shown below uses the symbol σ to represent the standard deviation. Note that 2σ represents the value that is two standard deviations above the mean; -2σ represents the value that is two standard deviations below the mean. The unshaded areas, from two standard deviations below the mean to two standard deviations above the mean, are considered the normal range. For example, the average (mean) 8-year-old boy has a height of about 50 inches, but any heights between approximately 45 inches and 55 inches are considered normal for 8-year-old boys. Refer to the chart below to answer the following questions.



a) What happens to the standard deviation for weights of boys as the age of boys increases? What is the significance of this fact?
b) At age 16, what is the mean weight, in pounds, of boys?
c) What is the approximate standard deviation of boys' weights at age 16?
d) Find the mean weight and normal range for boys at age 13.
e) Find the mean height and normal range for boys at age 13.
f) Assuming that this chart was constructed so that approximately 95% of all boys are always in the normal range, determine what percentage of boys are not in the normal range.

Challenge Problems/Group Activities

32. ***Athletes' Salaries*** The following table lists the 10 highest paid athletes in Major League Baseball and in the National Football League.

Major League Baseball
(2003 Season)

Player	Salary (millions)
1. Alex Rodriguez	$22
2. Manny Ramirez	20
3. Carlos Delgado	18.7
4. Mo Vaughn	17.2
5. Sammy Sosa	16
6. Kevin Brown	15.7
7. Shawn Green	15.7
8. Derek Jeter	15.6
9. Mike Piazza	15.6
10. Barry Bonds	15.5

Source: Major League Players Association.

National Football League
(2002 Season)

Player	Salary (millions)
1. Donovan McNabb	$15.4
2. Curtis Martin	13.3
3. Larry Allen	13.0
4. David Carr	12.0
5. Rod Smith	11.7
6. Jeff Garcia	11.7
7. Michael Strahan	11.4
8. Aaron Glenn	11.3
9. Tarik Glenn	11.3
10. Ray Lewis	10.5

Source: National Football League Players Association.

a) Without doing any calculations, which do you believe is greater, the mean salary of the 10 baseball players or the mean salary of the 10 football players? Explain.
b) Without doing any calculations, which do you believe is greater, the standard deviation of the salary of the 10 baseball players or the standard deviation of the salary of the 10 football players? Explain.
c) Compute the mean salary of the 10 baseball players and the mean salary of the 10 football players and determine whether your answer in part (a) was correct.
d) Compute the standard deviation of the salary of the 10 baseball players and the standard deviation of the salary of the 10 football players and determine whether your answer in part (b) is correct. Round each mean to the nearest tenth to determine the standard deviation.

33. ***Oil Change*** Jiffy Lube has franchises in two different parts of the city. The number of oil changes made daily, for 25 days, is given below.

East Store					West Store				
33	59	27	30	42	38	46	38	38	30
19	42	25	22	32	38	38	37	39	31
43	27	57	37	52	39	36	40	37	47
40	67	38	44	43	30	34	42	45	29
15	31	49	41	35	31	46	28	45	48

a) Construct a frequency distribution for each store with a first class of 15–20.
b) Draw a histogram for each store.
c) Using the histogram, determine which store appears to have a greater mean, or do the means appear about the same? Explain.
d) Using the histogram, determine which store appears to have the greater standard deviation? Explain.
e) Calculate the mean for each store and determine whether your answer in part (c) was correct.
f) Calculate the standard deviation for each store and determine whether your answer in part (d) was correct.

Recreational Mathematics

34. Calculate the range and standard deviation of your exam grades in this mathematics course. Round the mean to the nearest tenth to calculate the standard deviation.

35. Construct a set of 5 pieces of data with a mean, median, mode, and midrange of 6 and a standard deviation of 0.

Internet/Research Problem

36. Use a calculator with statistical function keys to find the mean and standard deviation of the salaries of the 10 Major League Baseball players and the 10 National Football League players in Exercise 32.

13.7 THE NORMAL CURVE

When examining data using a histogram, we can refer to the overall appearance of the histogram as the *shape* of the distribution of the data. Certain shapes of distributions of data are more common than others. In this section, we will illustrate and discuss a few of the more common ones. In each case, the vertical scale is the frequency and the horizontal scale is the observed values.

In a ***rectangular distribution*** (Fig. 13.17), all the observed values occur with the same frequency. If a die is rolled many times, we would expect the numbers 1–6 to occur with about the same frequency. The distribution representing the outcomes of the die is rectangular.

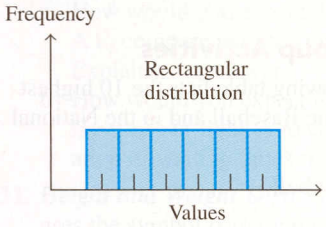


Figure 13.17

In ***J-shaped distributions***, the frequency is either constantly increasing (Fig. 13.18a) or constantly decreasing (Fig. 13.18b). The number of hours studied per week by students may have a distribution like that in Fig. 13.18(b). The bars might represent (from left to right) 0–5 hours, 6–10 hours, 11–15 hours, and so on.

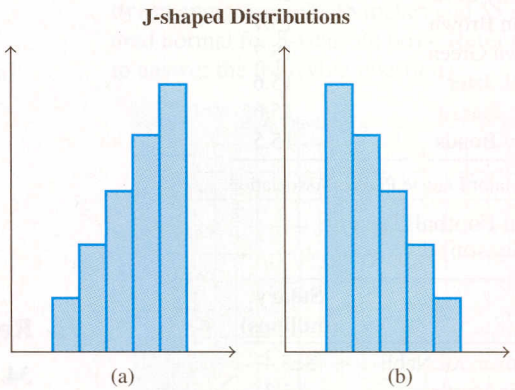


Figure 13.18

A ***bimodal distribution*** (Fig. 13.19) is one in which two nonadjacent values occur more frequently than any other values in a set of data. For example, if an equal number of men and women were weighed, the distribution of their weights would probably be bimodal, with one mode for the women's weights and the second for the men's weights. For a distribution to be considered bimodal, both modes need not have the same frequency but they must both have a frequency greater than the frequency of each of the other values in the distribution.

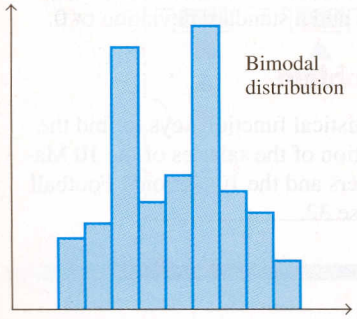


Figure 13.19

The life expectancy of light bulbs has a bimodal distribution: a small peak very near 0 hours of life, resulting from the bulbs that burned out very quickly because of a manufacturing defect, and a much broader peak representing the nondefective bulbs. A bimodal frequency distribution generally means that you are dealing with two distinct populations, in this case, defective and nondefective bulbs.

Another distribution, called a ***skewed distribution***, has more of a "tail" on one side than the other. A skewed distribution with a tail on the right (Fig. 13.20a) is said to be skewed to the right. If the tail is on the left (Fig. 13.20b), the distribution is referred to as skewed to the left.

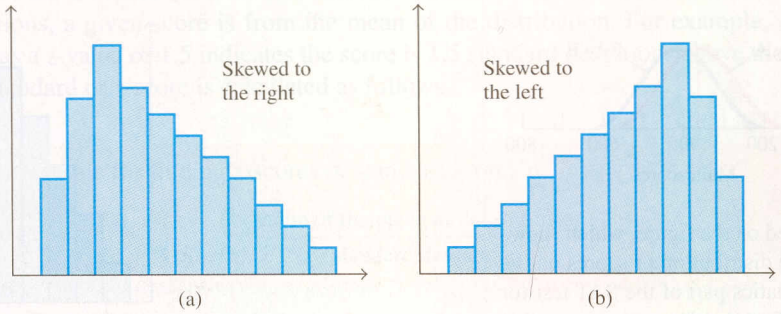


Figure 13.20

The number of children per family might be a distribution skewed to the right. Some families have no children, more families may have one child, the greatest percentage may have two children, fewer may have three children, still fewer may have four children, and so on.

Since few families have high incomes, distributions of family incomes might be skewed to the right.

Smoothing the histograms of the skewed distributions shown in Fig. 13.20 to form curves gives the curves illustrated in Fig. 13.21.

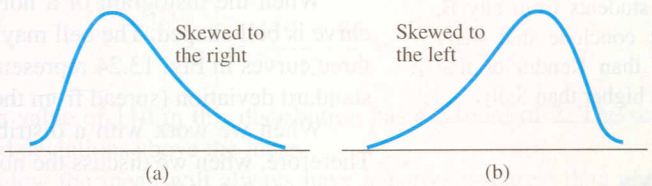


Figure 13.21

In Fig. 13.21(a), the greatest frequency appears on the left side of the curve, and the frequency decreases from left to right. Since the mode is the value with the greatest frequency, the mode would appear on the left side of the curve.

Every value in the set of data is considered in determining the mean. The values on the far right side of the curve in Fig. 13.21(a) would tend to increase the value of the mean. Thus, the value of the mean would be farther to the right than the mode. The median would be between the mode and the mean. The relationship between the mean, median, and mode for curves that are skewed to the right and left is given in Fig. 13.22.

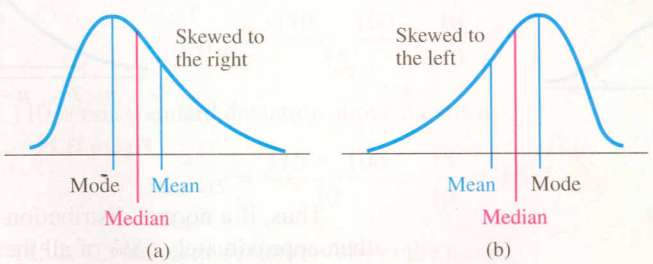
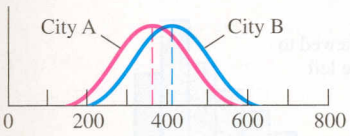


Figure 13.22

DID YOU KNOW

What Conclusions Can You Draw?



Based on the figure, which shows the distribution of scores on the mathematics part of the SAT test for two different cities, can we say that any given person selected at random from city B has outperformed any given person selected at random from city A? Both distributions appear normal, and the mean of city A is slightly smaller than the mean of city B. Consider, however, two randomly selected students who took this test: Sally from city A and Kendra from city B. The graph shows that many students in city A outperformed students from city B, so we cannot conclude that Sally scored higher than Kendra or that Kendra scored higher than Sally.

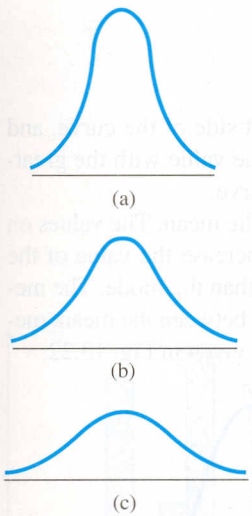


Figure 13.24

Each of these distributions is useful in describing sets of data. However, the most important distribution is the ***normal*** or ***Gaussian distribution***, named for the German mathematician Carl Friedrich Gauss. The histogram of a normal distribution is illustrated in Fig. 13.23.

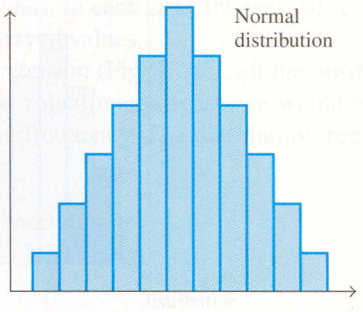


Figure 13.23

The normal distribution is important because many sets of data are normally distributed, or they closely resemble a normal distribution. Such distributions include intelligence quotients, heights and weights of males, heights and weights of females, lengths of full-grown boa constrictors, weights of watermelons, wearout mileage of automobile brakes, and life spans of refrigerators, to name just a few.

The normal distribution is symmetric about the mean. If you were to fold the histogram down the middle, the left side would fit the right side exactly. **In a normal distribution, the mean, median, and mode all have the same value.**

When the histogram of a normal distribution is smoothed to form a curve, the curve is bell-shaped. The bell may be high and narrow or short and wide. Each of the three curves in Fig. 13.24 represents a normal curve. Curve 13.24(a) has the smallest standard deviation (spread from the mean); curve 13.24(c) has the largest.

When we work with a distribution, we are working with an entire population. Therefore, when we discuss the normal distribution, we use μ for the mean and σ for the standard deviation.

Since the curve is symmetric, 50% of the data always falls above (to the right of) the mean and 50% of the data falls below (to the left of) the mean. In addition, every normal distribution has approximately 68% of the data between the value that is one standard deviation below the mean, $\mu - 1\sigma$, and the value that is one standard deviation above the mean, $\mu + 1\sigma$; see Fig. 13.25. Approximately 95% of the data falls between the value that is two standard deviations below the mean, $\mu - 2\sigma$, and the value that is two standard deviations above the mean, $\mu + 2\sigma$.

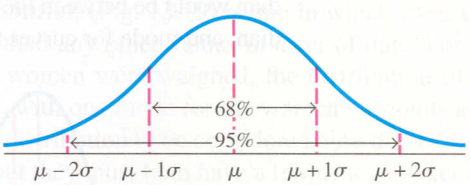


Figure 13.25

Thus, if a normal distribution has a mean of 100 and a standard deviation of 10, then approximately 68% of all the data falls between 100 − 10 and 100 + 10, or be-

PROFILE IN MATHEMATICS

DAVID BLACKWELL (1919–)

David H. Blackwell, (1919–), professor of statistics, is the author of more than 90 publications on statistics, probability, game theory, set theory, dynamic programming, and information theory. Blackwell, past president of the American Statistical Society, was the first African-American elected to the National Academy of Sciences.

His first interests in mathematics were in Geometry as a schoolboy growing up in southern Illinois. When he enrolled at the University of Illinois at the age of 16, he planned to be an elementary school teacher. When he received his Ph.D. in mathematics from the University of Illinois in 1941, he was just the sixth African-American to receive a doctorate in mathematics in the country.

After spending 2 years as a postdoctoral fellow at the Institute for Advanced Study at Princeton, he taught mathematics for 10 years at Howard University. While at Howard, Blackwell built a strong national reputation as a gifted teacher and creative researcher.

In 1954, he joined the Department of Statistics at the University of California, Berkeley. Blackwell, who has taught a wide variety of mathematics courses, says, "Basically, I'm not interested in doing research and I never have been. I'm interested in understanding, which is quite a different thing."

tween 90 and 110. Approximately 95% of the data falls between 100 − 20 and 100 + 20, or between 80 and 120. In fact, given any normal distribution with a known standard deviation and mean, it is possible through the use of Table 13.7 page 801 (the z-table) to determine the percent of data between any two given values.

We use ***z-scores*** (or ***standard scores***) to determine how far, in terms of standard deviations, a given score is from the mean of the distribution. For example, a score that has a z-value of 1.5 indicates the score is 1.5 standard deviations above the mean. The standard or z-score is calculated as follows.

The formula for finding z-scores or standard scores is

$$z = \frac{\text{value of the piece of data} - \text{mean}}{\text{standard deviation}}$$

If we let x represent the value of the given piece of data, μ represent the mean, and σ represent the standard deviation, we can symbolize the z-score formula as

$$z = \frac{x - \mu}{\sigma}.$$

In this book, the notation z_x represents the z-score, or standard score, of the value x. For example, if a normal distribution has a mean of 86 with a standard deviation of 12, a score of 110 has a standard or z-score of

$$z_{110} = \frac{110 - 86}{12} = \frac{24}{12} = 2$$

Therefore, a value of 110 in this distribution has a z-score of 2. The score of 110 is two standard deviations above the mean.

Data below the mean will always have negative z-scores; data above the mean will always have positive z-scores. The mean will always have a z-score of 0.

EXAMPLE 1 *Finding z-Scores*

A normal distribution has a mean of 100 and a standard deviation of 10. Find z-scores for the following values.

a) 110 b) 115 c) 100 d) 84

SOLUTION:

a)
$$z = \frac{\text{value} - \text{mean}}{\text{standard deviation}}$$
$$z_{110} = \frac{110 - 100}{10} = \frac{10}{10} = 1$$

A score of 110 is one standard deviation above the mean.

b)
$$z_{115} = \frac{115 - 100}{10} = \frac{15}{10} = 1.5$$

A score of 115 is 1.5 standard deviations above the mean.

DID YOU KNOW

Six Sigma

Many companies use a process called Six Sigma, a quality control strategy, to help the company improve quality and reduce errors. Six Sigma refers to an interval in a normal distribution from six standard deviations below the mean to six standard deviations above the mean. As 99.9997% of a normal distribution is within six standard deviations of the mean, Six Sigma means the company's goal is to produce error-free products 99.9997% of the time. Companies such as General Electric (GE), Whirlpool, and Motorola have all reported success after implementing Six Sigma. GE's former CEO, Jack Welch, a major advocate for Six Sigma, reported that Six Sigma saved the company more than $2 billion in 1999.

c) $z_{100} = \dfrac{100 - 100}{10} = \dfrac{0}{10} = 0$

The mean always has a z-score of 0.

d) $z_{84} = \dfrac{84 - 100}{10} = \dfrac{-16}{10} = -1.6$

A score of 84 is 1.6 standard deviations below the mean. ▲

Let's now consider finding areas under the normal curve. The total area under the normal curve is 1.00. Table 13.7 on page 801 will be used to determine the area under the normal curve between any two given points (the values in the table have been rounded). **Table 13.7 gives the area under the normal curve from the mean (a z-value of 0) to a z-value to the right of the mean.**

For example, between the mean and $z = 2.00$, the table shows a value of 0.477. Thus, there is 0.477 of the total area under the curve between the mean and $z = 2.00$, see Fig. 13.26. To change this area of 0.477 to a percent, simply multiply by 100%: $0.477 \times 100\%$ is 47.7%. Thus, 47.7% of all scores will be between the mean and the score that is two standard deviations above the mean.

When you are finding the area under the normal curve, it is often helpful to draw a picture such as the one in Fig. 13.26, indicating the area or percent to be found.

The normal curve is symmetric about the mean. Thus, the same percent of data is between the mean and a positive z-score as between the mean and the corresponding negative z-score. For example, there is the same area under the normal curve between a z of 1.60 and the mean as between a z of -1.60 and the mean. By looking up a z-score of 1.60 in Table 13.7, we see that both have an area of 0.445 (Fig. 13.27). Since an area of 0.445 corresponds to 44.5%, we can reason that 44.5% + 44.5% or 89.0% of the data is between z-scores of -1.60 and 1.60.

You now have the necessary knowledge to find the percent of data between any two values in a normal distribution.

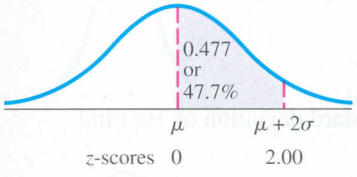


Figure 13.26

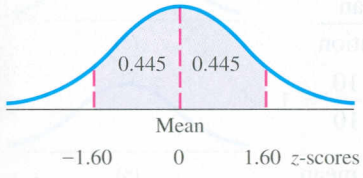
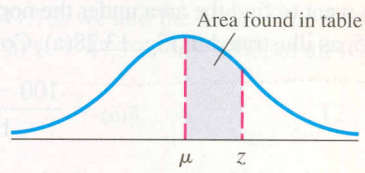


Figure 13.27

To Find the Percent of Data Between any Two Values:

1. Draw a diagram of the normal curve, indicating the area or percent to be determined.
2. Use the formula $z = \dfrac{x - \mu}{\sigma}$ to convert the given values to z-scores. Indicate these z-scores on the diagram.
3. Look up the percent that corresponds to each z-score in Table 13.7.
4. **a)** When finding the percent of data between two z-scores on the opposite side of the mean (when one z-score is positive and the other is negative), you find the sum of the individual percents.
   b) When finding the percent of data between two z-scores on the same side of the mean (when both z-scores are positive or both are negative), subtract the smaller percent from the larger percent.
   c) When finding the percent of data to the right of a positive z-score or to the left of a negative z-score, subtract the percent of data between 0 and z from 50%.
   d) When finding the percent of data to the left of a positive z-score or to the right of a negative z-score, add the percent of data between 0 and z to 50%.

TABLE 13.7 Areas under the Standard Normal Curve (the z-table)



The column under A gives the area under the entire curve that is between $z = 0$ (or the mean) and a positive value of z

z	A	z	A	z	A	z	A	z	A	z	A	z	A	z	A	z	A
.00	.000	.37	.144	.74	.270	1.11	.367	1.48	.431	1.85	.468	2.22	.487	2.59	.495	2.96	.499
.01	.004	.38	.148	.75	.273	1.12	.369	1.49	.432	1.86	.469	2.23	.487	2.60	.495	2.97	.499
.02	.008	.39	.152	.76	.276	1.13	.371	1.50	.433	1.87	.469	2.24	.488	2.61	.496	2.98	.499
.03	.012	.40	.155	.77	.279	1.14	.373	1.51	.435	1.88	.470	2.25	.488	2.62	.496	2.99	.499
.04	.016	.41	.159	.78	.282	1.15	.375	1.52	.436	1.89	.471	2.26	.488	2.63	.496	3.00	.499
.05	.020	.42	.163	.79	.285	1.16	.377	1.53	.437	1.90	.471	2.27	.488	2.64	.496	3.01	.499
.06	.024	.43	.166	.80	.288	1.17	.379	1.54	.438	1.91	.472	2.28	.489	2.65	.496	3.02	.499
.07	.028	.44	.170	.81	.291	1.18	.381	1.55	.439	1.92	.473	2.29	.489	2.66	.496	3.03	.499
.08	.032	.45	.174	.82	.294	1.19	.383	1.56	.441	1.93	.473	2.30	.489	2.67	.496	3.04	.499
.09	.036	.46	.177	.83	.297	1.20	.385	1.57	.442	1.94	.474	2.31	.490	2.68	.496	3.05	.499
.10	.040	.47	.181	.84	.300	1.21	.387	1.58	.443	1.95	.474	2.32	.490	2.69	.496	3.06	.499
.11	.044	.48	.184	.85	.302	1.22	.389	1.59	.444	1.96	.475	2.33	.490	2.70	.497	3.07	.499
.12	.048	.49	.188	.86	.305	1.23	.391	1.60	.445	1.97	.476	2.34	.490	2.71	.497	3.08	.499
.13	.052	.50	.192	.87	.308	1.24	.393	1.61	.446	1.98	.476	2.35	.491	2.72	.497	3.09	.499
.14	.056	.51	.195	.88	.311	1.25	.394	1.62	.447	1.99	.477	2.36	.491	2.73	.497	3.10	.499
.15	.060	.52	.199	.89	.313	1.26	.396	1.63	.449	2.00	.477	2.37	.491	2.74	.497	3.11	.499
.16	.064	.53	.202	.90	.316	1.27	.398	1.64	.450	2.01	.478	2.38	.491	2.75	.497	3.12	.499
.17	.068	.54	.205	.91	.319	1.28	.400	1.65	.451	2.02	.478	2.39	.492	2.76	.497	3.13	.499
.18	.071	.55	.209	.92	.321	1.29	.402	1.66	.452	2.03	.479	2.40	.492	2.77	.497	3.14	.499
.19	.075	.56	.212	.93	.324	1.30	.403	1.67	.453	2.04	.479	2.41	.492	2.78	.497	3.15	.499
.20	.079	.57	.216	.94	.326	1.31	.405	1.68	.454	2.05	.480	2.42	.492	2.79	.497	3.16	.499
.21	.083	.58	.219	.95	.329	1.32	.407	1.69	.455	2.06	.480	2.43	.493	2.80	.497	3.17	.499
.22	.087	.59	.222	.96	.332	1.33	.408	1.70	.455	2.07	.481	2.44	.493	2.81	.498	3.18	.499
.23	.091	.60	.226	.97	.334	1.34	.410	1.71	.456	2.08	.481	2.45	.493	2.82	.498	3.19	.499
.24	.095	.61	.229	.98	.337	1.35	.412	1.72	.457	2.09	.482	2.46	.493	2.83	.498	3.20	.499
.25	.099	.62	.232	.99	.339	1.36	.413	1.73	.458	2.10	.482	2.47	.493	2.84	.498	3.21	.499
.26	.103	.63	.236	1.00	.341	1.37	.415	1.74	.459	2.11	.483	2.48	.493	2.85	.498	3.22	.499
.27	.106	.64	.239	1.01	.344	1.38	.416	1.75	.460	2.12	.483	2.49	.494	2.86	.498	3.23	.499
.28	.110	.65	.242	1.02	.346	1.39	.418	1.76	.461	2.13	.483	2.50	.494	2.87	.498	3.24	.499
.29	.114	.66	.245	1.03	.349	1.40	.419	1.77	.462	2.14	.484	2.51	.494	2.88	.498	3.25	.499
.30	.118	.67	.249	1.04	.351	1.41	.421	1.78	.463	2.15	.484	2.52	.494	2.89	.498	3.26	.499
.31	.122	.68	.252	1.05	.353	1.42	.422	1.79	.463	2.16	.485	2.53	.494	2.90	.498	3.27	.500
.32	.126	.69	.255	1.06	.355	1.43	.424	1.80	.464	2.17	.485	2.54	.495	2.91	.498	3.28	.500
.33	.129	.70	.258	1.07	.358	1.44	.425	1.81	.465	2.18	.485	2.55	.495	2.92	.498	3.29	.500
.34	.133	.71	.261	1.08	.360	1.45	.427	1.82	.466	2.19	.486	2.56	.495	2.93	.498	3.30	.500
.35	.137	.72	.264	1.09	.362	1.46	.428	1.83	.466	2.20	.486	2.57	.495	2.94	.498	3.31	.500
.36	.141	.73	.267	1.10	.364	1.47	.429	1.84	.467	2.21	.487	2.58	.495	2.95	.498	3.32	.500

EXAMPLE 2 *IQ Scores*

Intelligence quotients (IQs) are normally distributed with a mean of 100 and a standard deviation of 15. Find the percent of individuals with IQs

a) between 100 and 115.
b) between 70 and 100.
c) between 70 and 115.
d) between 115 and 130.
e) below 130.
f) above 122.5.

SOLUTION:

a) We want to find the area under the normal curve between the values of 100 and 115, as illustrated in Fig. 13.28(a). Converting 100 to a z-score yields a z-score of 0.

$$z_{100} = \frac{100 - 100}{15} = \frac{0}{15} = 0$$

Converting 115 to a z-score yields a z-score of 1.00.

$$z_{115} = \frac{115 - 100}{15} = \frac{15}{15} = 1.00$$

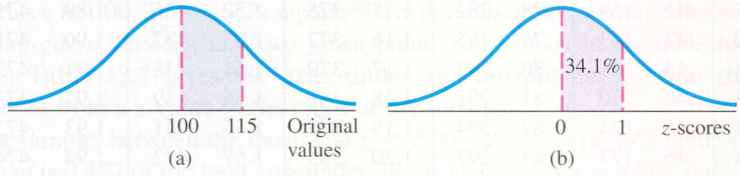


Figure 13.28

The percent of individuals with IQs between 100 and 115 is the same as the percent of data between z-scores of 0 and 1 (Fig. 13.28b).

From Table 13.7, we determine that 0.341 of the area, or 34.1% of all the data, is between z-scores of 0 and 1.00. Therefore, 34.1% of individuals have IQs between 100 and 115.

b) Begin by finding z-scores for 70 and 100.

$$z_{70} = \frac{70 - 100}{15} = \frac{-30}{15} = -2.00$$

$$z_{100} = 0 \text{ (from part a)}$$

The percent of data between scores of 70 and 100 is the same as the percent between $z = -2$ and $z = 0$ (Fig. 13.29). The percent of data between the mean and two standard deviations below the mean is the same as the percent of data between the mean and two standard deviations above the mean. From Table 13.7, we determine that 47.7% of the data is between $z = 0$ and $z = 2$. Thus, 47.7% of the data is also between $z = -2.00$ and $z = 0$. Therefore, 47.7% of all individuals have IQs between 70 and 100.

c) In parts (a) and (b), we determined that $z_{115} = 1.00$ and $z_{70} = -2.00$. Since the values are on opposite sides of the mean, the percent of data between the two values is found by adding the individual percents: 34.1% + 47.7% = 81.8% (Fig. 13.30). Thus, 81.8% of the IQs are between 70 and 115.

d) Begin by finding z-scores for 115 and 130.

$$z_{115} = 1.00 \text{ (from part a)}$$

$$z_{130} = \frac{130 - 100}{15} = \frac{30}{15} = 2.00$$

Since both values are on the same side of the mean (Fig. 13.31), the smaller percent must be subtracted from the larger percent to obtain the percent of data in the shaded area: 47.7% − 34.1% is 13.6%. Thus, 13.6% of all the individuals have IQs between 115 and 130.

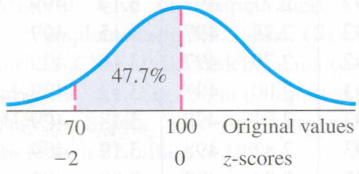


Figure 13.29

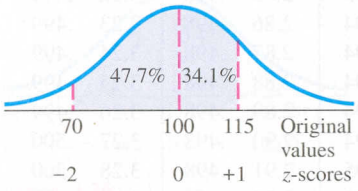


Figure 13.30

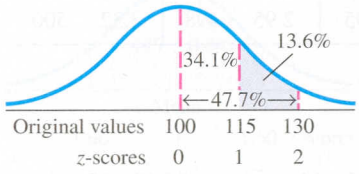


Figure 13.31

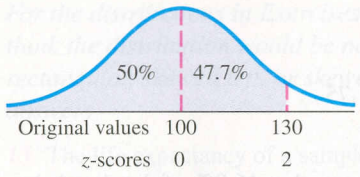


Figure 13.32

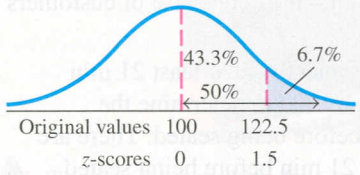


Figure 13.33

e) The percent of IQs below 130 is the same as the percent of data below a z-score of 2. Between $z = 2$ and the mean is 47.7% of the data (Fig. 13.32). To this 47.7% we add the 50% of the data below the mean to give 97.7%. Thus, 50% + 47.7%, or 97.7%, of all IQs are below 130.

f) $$z_{122.5} = \frac{122.5 - 100}{15} = \frac{22.5}{15} = 1.50$$

The percent of IQs above 122.5 is the same as the percent of data above $z = 1.5$ (Fig. 13.33). Fifty percent of the data is to the right of the mean. Since 43.3% of the data is between the mean and $z = 1.5$, 50% − 43.3%, or 6.7%, of the data is greater than $z = 1.5$. Thus, 6.7% of all IQs are greater than 122.5.

EXAMPLE 3 *Waiting Time at a Restaurant*

Assume that the waiting times for customers at a popular restaurant before being seated for lunch are normally distributed with a mean of 16 min and standard deviation of 4 min.

a) Find the percent of customers who wait for at least 16 min before being seated.

b) Find the percent of customers who wait between 12 and 24 min before being seated.

c) Find the percent of customers who wait at least 21 min before being seated.

d) Find the percent of customers who wait less than 9 min before being seated.

e) In a random sample of 500 customers, how many wait at least 21 min before being seated?

SOLUTION:

a) In a normal distribution, half the data are always above the mean. Since 16 min is the mean, half, or 50%, of customers wait at least 16 min before being seated.

b) Convert 12 min and 24 min to z-scores.

$$z_{12} = \frac{12 - 16}{4} = -1.00$$

$$z_{24} = \frac{24 - 16}{4} = 2.00$$

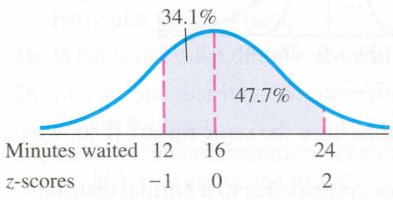


Figure 13.34

Now look up the areas in Table 13.7. The percent of customers who wait between 12 and 24 min before being seated is 34.1% + 47.7% or 81.8% (Fig. 13.34).

c) Convert 21 min to a z-score.

$$z_{21} = \frac{21 - 16}{4} = 1.25$$

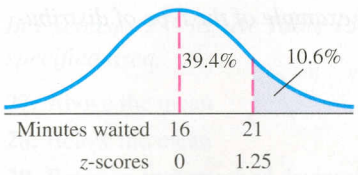


Figure 13.35

Look up the area in Table 13.7. Figure 13.35 shows 39.4% of the data is between the mean and $z = 1.25$. Therefore, the percent of data above $z = 1.25$ is 50% − 39.4% = 10.6%. Thus, 10.6% of customers wait at least 21 min before being seated.

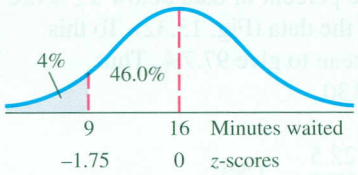


Figure 13.36

d) Convert 9 min to a z-score.

$$z_9 = \frac{9 - 16}{4} = -1.75$$

Look up the area in Table 13.7. Figure 13.36 shows that 46.0% of the data is between the mean and $z = -1.75$. The percent of data to the left of $z = -1.75$ is found by subtracting 46.0% from 50.0% to obtain 4.0%. Thus, 4% of customers wait less than 9 min before being seated.

e) In part (c), we determined that 10.6% of all customers wait at least 21 min before being seated. We now multiply 0.106 times 500 to determine the number of customers who wait at least 21 min before being seated. There are $0.106 \times 500 = 53$ customers who wait at least 21 min before being seated. ▲

TIMELY TIP Remember that area cannot be negative. A negative z-score indicates that the corresponding value in the original distribution is below the mean.

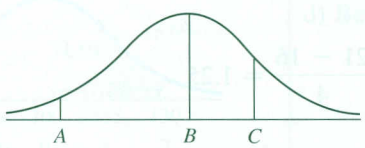
SECTION 13.7 EXERCISES

Concept/Writing Exercises

In Exercises 1–6, describe

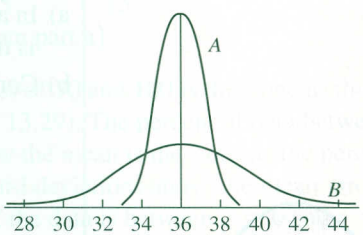
1. a rectangular distribution.
2. a J-shaped distribution.
3. a bimodal distribution.
4. a distribution that is skewed to the right.
5. a distribution that is skewed to the left.
6. a normal distribution.

7. Consider the following normal curve, representing a normal distribution, with points A, B, and C. One of these points corresponds to μ, one point corresponds to $\mu + \sigma$, and one point corresponds to $\mu - 2\sigma$.



a) Which point corresponds to μ?
b) Which point corresponds to $\mu + \sigma$?
c) Which point corresponds to $\mu - 2\sigma$?

8. Consider the following normal curves.



a) Do these distributions have the same mean? If so, what is the mean?
b) One of these curves corresponds to a normal distribution with $\sigma = 1$. The other curve corresponds to a normal distribution with $\sigma = 3$. Which curve, A or B, has $\sigma = 3$? Explain.

In Exercises 9–12, give an example of the type of distribution.

9. Rectangular
10. Skewed
11. J-shaped
12. Bimodal

For the distributions in Exercises 13–16, state whether you think the distribution would be normal, J-shaped, bimodal, rectangular, skewed left, or skewed right. Explain your answers.

13. The life expectancy of a sample of microwave ovens

14. The numbers resulting from tossing a die many times
15. The salaries of teachers at Roosevelt High School, where there are many newly hired teachers
16. The heights of a sample of high school seniors, where there are an equal number of males and females
17. In a distribution that is skewed to the right, which has the greatest value: the mean, median, or mode? Which has the smallest value? Explain.
18. In a distribution skewed to the left, which has the greatest value: the mean, median, or mode? Which has the smallest value? Explain.
19. List three populations other than those given in the text that may be normally distributed.
20. List three populations other than those given in the text that may not be normally distributed.
21. In a normal distribution, what is the relationship between the mean, median, and mode?
22. What does the z, or standard score, measure?
23. When will a z-score be negative?
24. Explain in your own words how to find the z-score of a particular piece of data.
25. What is the value of the z-score of the mean of a set of data?
26. In a normal distribution, approximately what percent of the data is between
    a) one standard deviation below the mean to one standard deviation above the mean?
    b) two standard deviations below the mean to two standard deviations above the mean?

Practice the Skills

In Exercises 27–38, use Table 13.7 on page 801 to find the specified area.

27. Above the mean
28. Below the mean
29. Between two standard deviations below the mean and one standard deviation above the mean
30. Between 1.10 and 1.70 standard deviations above the mean
31. To the right of $z = 1.82$
32. To the left of $z = 1.19$
33. To the left of $z = -1.78$
34. To the right of $z = -1.78$
35. To the right of $z = 2.08$
36. To the left of $z = 1.96$
37. To the left of $z = -1.62$
38. To the left of $z = -0.90$

In Exercises 39–48, use Table 13.7 to determine the percent of data specified.

39. Between $z = 0$ and $z = 0.71$
40. Between $z = -0.15$ and $z = -0.82$
41. Between $z = -1.34$ and $z = 2.24$
42. Less than $z = -1.90$
43. Greater than $z = -1.90$
44. Greater than $z = 2.66$
45. Less than $z = 1.96$
46. Between $z = 0.72$ and $z = 2.14$
47. Between $z = -1.53$ and $z = -1.82$
48. Between $z = -2.15$ and $z = 3.31$

Problem Solving

Fitness Test Scores *In Exercises 49 and 50, suppose that the results on a fitness test are normally distributed. The z-scores for some participants are shown below.*

Jake	1.3	Marie	0.0	Justin	−1.9	Kevin	0.0
Sarah	1.7	Omar	−2.1	Carol	0.8	Kim	−1.2

49. a) Which of these participants scored above the mean?
    b) Which of these participants scored at the mean?
    c) Which of these participants scored below the mean?
50. a) Which participant had the highest score?
    b) Which participant had the lowest score?

Hours Worked by College Students *In Exercises 51–54, assume that the number of hours college students spend working per week is normally distributed with a mean of 18 hours and standard deviation of 4 hours.*

51. Find the percent of college students who work at least 18 hours per week.
52. Find the percent of college students who work between 14 and 26 hours per week.

53. Determine the percent of college students who work at least 23 hours per week.

54. In a random sample of 500 college students, how many work at least 23 hours per week?

Watching Television *In Exercises 55–60, assume that the amount of time children spend watching television per year is normally distributed with a mean of 1600 hours and a standard deviation of 100 hours.*

55. What percent of children watch television less than 1650 hours per year?

56. What percent of children watch television more than 1750 hours per year?

57. What percent of children watch television between 1650 and 1750 hours per year?

58. What percent of children watch television less than 1400 hours per year?

59. What percent of children watch television between 1500 and 1625 hours per year?

60. What percent of children watch television more than 1480 hours per year?

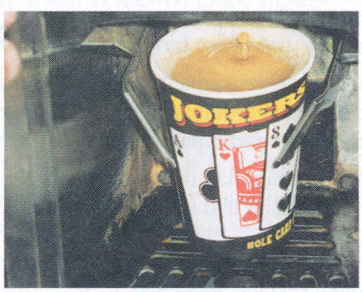
Vending Machine *In Exercises 61–64, a vending machine is designed to dispense a mean of 7.6 oz of coffee into an 8 oz cup. If the standard deviation of the amount of coffee dispensed is 0.4 oz and the amount is normally distributed, find the percent of times the machine will*

61. dispense from 7.4 oz to 7.7 oz.

62. dispense less than 7.0 oz.

63. dispense less than 7.7 oz.

64. result in the cup overflowing (therefore dispense more than 8 oz).



Cholesterol Levels *In Exercises 65–70, assume that the cholesterol levels for females are normally distributed with a mean of 206 and a standard deviation of 12.*

65. What percent of females have a cholesterol level greater than 206?

66. What percent of females have a cholesterol level between 197 and 215?

67. What percent of females have a cholesterol level less than 191?

68. What percent of females have a cholesterol level greater than 224?

69. If 200 women are selected at random, how many will have a cholesterol level less than 191?

70. If 200 women are selected at random, how many will have a cholesterol level greater than 224?

Tire Mileage *In Exercises 71–74, the wearout mileage of a certain tire is normally distributed with a mean of 35,000 miles and standard deviation of 2500 miles.*

71. Determine the percent of tires that will last between 30,750 miles and 38,300 miles.

72. Determine the percent of tires that will last at least 39,000 miles.

73. If the manufacturer guarantees the tires to last at least 30,750 miles, what percent of tires will fail to live up to the guarantee?

74. If 200,000 tires are produced, how many will last at least 39,000 miles?

Ages of Children at Day Care *In Exercises 75–80, assume that the ages of children at Happy Times Day Care are normally distributed with a mean of 3.7 years and a standard deviation of 1.2 years. Find the percent of children at Happy Times Day Care who are*

75. Older than 3.1 years.

76. Between 2.5 and 4.3 years.

77. Older than 6.7 years.

78. Younger than 6.7 years.

79. If 120 children are enrolled at Happy Times Day Care, how many of them are older than 3.1 years?

80. If 120 children are enrolled at Happy Times Day Care, how many of them are between 2.5 and 4.3 years?

81. ***Weight Loss*** A weight-loss clinic guarantees that its new customers will lose at least 5 lb by the end of their first month of participation or their money will be refunded. If the loss of weight of customers at the end of their first month is normally distributed, with a mean of 6.7 lb and a standard deviation of 0.81 lb, find the percent of customers who will be able to claim a refund.

82. ***Appliance Warranty*** The warranty on the motor of a dishwasher is 8 yr. If the breakdown times of this motor are normally distributed, with a mean of 10.2 yr and a standard deviation of 1.8 yr, find the percent of motors that can be expected to require repair or replacement under warranty.

83. ***Coffee Machine*** A vending machine that dispenses coffee does not appear to be working correctly. The machine rarely gives the proper amount of coffee. Some of the time the cup is underfilled, and some of the time the cup overflows. Does this variation indicate that the mean number of ounces dispensed has to be adjusted or that the standard deviation of the amount of coffee dispensed by the machine is too large? Explain your answer.

84. ***Grading on a Normal Curve*** Mr. Sanderson marks his class on a normal curve. Those with z-scores above 1.8 will receive an A, those between 1.8 and 1.1 will receive a B, those between 1.1 and -1.2 will receive a C, those between -1.2 and -1.9 will receive a D, and those under -1.9 will receive an F. Find the percent of grades that will be A, B, C, D, and F.

Challenge Problems/Group Activities

85. ***Salesperson Promotion*** The owner at Kim's Home Interiors is reviewing the sales records of two managers who are up for promotion, Katie and Stella, who work in different stores. At Katie's store, the mean sales have been \$23,200 per month, with a standard deviation of \$2170. At Stella's store, the mean sales have been \$25,600 per month, with a standard deviation of \$2300. Last month Katie's store sales were \$28,408 and Stella's store sales were \$29,510. At both stores, the distribution of monthly sales is normal.

a) Convert last month's sales for Katie's store and for Stella's store to z-scores.

b) If one of the two were to be promoted based solely on the increase in sales last month, who should be promoted? Explain.

86. ***Chebyshev's Theorem*** How can you determine whether a distribution is approximately normal? A statistical theorem called ***Chebyshev's theorem*** states that the *minimum percent* of data between plus and minus K standard deviations from the mean $(K > 1)$ in *any distribution* can be found by the formula

$$\text{Minimum percent} = 1 - \frac{1}{K^2}$$

Thus, for example, between ± 2 standard deviations from the mean, there will always be a minimum of 75% of data. This minimum percent is true for any distribution. For $K = 2$,

$$\text{Minimum percent} = 1 - \frac{1}{2^2} = 1 - \frac{1}{4} = \frac{3}{4}, \quad \text{or} \quad 75\%$$

Likewise, between ± 3 standard deviations from the mean, there will always be a minimum of 89% of the data. For $K = 3$,

$$\text{Minimum percent} = 1 - \frac{1}{3^2} = 1 - \frac{1}{9} = \frac{8}{9}, \quad \text{or} \quad 89\%$$

The following table lists the minimum percent of data in *any distribution* and the actual percent of data in *the normal distribution* between ± 1.1, ± 1.5, ± 2.0, and ± 2.5 standard deviations from the mean. The minimum percents of data in any distribution were calculated by using Chebyshev's theorem. The actual percents of data for the normal distribution were calculated by using the area given in the standard normal, or z, table.

Given distribution	$K = 1.1$	$K = 1.5$	$K = 2$	$K = 2.5$
Minimum (for any distribution)	17.4%	55.6%	75%	84%
Normal distribution	72.8%	86.6%	95.4%	99.8%

The third row of the chart has been left blank for you to fill in the percents when you reach part (e).

Consider the following 30 pieces of data obtained from a quiz.

1, 1, 1, 1, 2, 2, 2, 2, 3, 3, 4, 4, 4, 5, 6,
6, 6, 7, 7, 7, 7, 8, 8, 8, 8, 9, 9, 9, 10, 10

a) Find the mean of the set of scores.

b) Find the standard deviation of the set of scores.

c) Determine the values that correspond to 1.1, 1.5, 2, and 2.5 standard deviations above the mean. (For example, the value that corresponds to 1.5 standard deviations above the mean is $\mu + 1.5\sigma$.)

Then determine the values that correspond to 1.1, 1.5, 2, and 2.5 standard deviations below the mean. (For example, the value that corresponds to 1.5 standard deviations below the mean is $\mu - 1.5\sigma$.)

d) By observing the 30 pieces of data, determine the actual percent of quiz scores between

± 1.1 standard deviations from the mean.

± 1.5 standard deviations from the mean.

± 2 standard deviations from the mean.

± 2.5 standard deviations from the mean.

e) Place the percents found in part (d) in the third row of the chart.

f) Compare the percents in the third row of the chart with the minimum percents in the first row and the normal

percents in the second row, and then make a judgment as to whether this set of 30 scores is approximately normally distributed. Explain your answer.

87. *Using Data from Your Class* Obtain a set of test scores from your teacher.
   a) Find the mean, median, mode, and midrange of the test scores.
   b) Find the range and standard deviation of the set of scores. (You may round the mean to the nearest tenth when finding the standard deviation.)
   c) Construct a frequency distribution of the set of scores. Select your first class so that there will be between 5 and 12 classes.
   d) Construct a histogram and frequency polygon of the frequency distribution in part (c).
   e) Does the histogram in part (d) appear to represent a normal distribution? Explain.
   f) Use the procedure explained in Exercise 86 to determine whether the set of scores approximates a normal distribution. Explain.

88. Find a value of z such that $z \geq 0$ and 47.5% of the standard normal curve lies between 0 and the z-value.

89. Find a value of z such that $z \leq 0$ and 38.1% of the standard normal curve lies between 0 and the z-value.

Recreational Mathematics

90. Ask your instructor for the class mean and class standard deviation for one of the exams taken by your class. For that exam, calculate the z-score for your exam grade. How many standard deviations is your exam grade away from the mean?

91. If the mean score on a math quiz is 12.0 and 77% of the students in your class scored between 9.6 and 14.4, determine the standard deviation of the quiz scores.

Internet/Research Activity

92. In this project, you actually become the statistician.
   a) Select a project of interest to you in which data must be collected.
   b) Write a proposal and submit it to your instructor for approval. In the proposal, discuss the aims of your project and how you plan to gather the data to make your sample unbiased.
   c) After your proposal has been approved, gather 50 pieces of data by the method you proposed.
   d) Rank the data from smallest to largest.
   e) Compute the mean, median, mode, and midrange.
   f) Determine the range and standard deviation of the data. You may round the mean to the nearest tenth when computing the standard deviation.
   g) Construct a frequency distribution, histogram, frequency polygon, and stem-and-leaf display of your data. Select your first class so that there will be between 5 and 12 classes. Be sure to label your histogram and frequency polygon.
   h) Does your distribution appear to be normal? Explain your answer. Does it appear to be another type of distribution discussed? Explain.
   i) Determine whether your distribution is approximately normal by using the technique discussed in Exercise 86.

13.8 LINEAR CORRELATION AND REGRESSION

In this section, we discuss two important statistical topics: correlation and regression. *Correlation* is used to determine whether there is a relationship between two quantities and, if so, how strong that relationship is. *Regression* is used to determine the equation that relates the two quantities. Although there are other types of correlation and regression, in this section we discuss only linear correlation and linear regression. We begin by discussing linear correlation.

Linear Correlation

Do you believe that there is a relationship between

a) the time a person studied for an exam and the exam grade received?
b) the age of a car and the value of the car?
c) the height and weight of adult males?
d) a person's IQ and income?

Correlation is used to answer questions of this type. The *linear correlation coefficient*, r, is a unitless measure that describes the strength of the linear relationship between two variables. A positive value of r, or a positive correlation, means that as one variable increases, the other variable also increases. A negative value of r, or a negative

correlation, means that as one variable increases, the other variable decreases. The correlation coefficient, r, will always be a value between -1 and 1 inclusive. A value of 1 indicates the strongest possible positive correlation, a value of -1 indicates the strongest possible negative correlation, and a value of 0 indicates no correlation (Fig. 13.37).

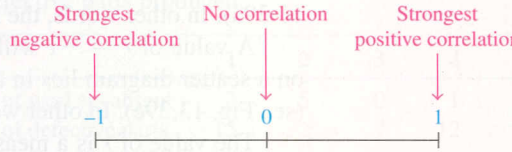


Figure 13.37

A visual aid used with correlation is the ***scatter diagram***, a plot of data points. To help understand how to construct a scatter diagram, consider the following data from Egan Electronics. During a 6-day period, Egan Electronics kept daily records of the number of assembly line workers absent and the number of defective parts produced. The information is provided in the following chart.

Day	1	2	3	4	5	6
Number of workers absent	3	5	0	1	2	6
Number of defective parts	15	22	7	12	20	30

For each of the 6 days, two pieces of data are provided: number of workers absent and number of defective parts. We call the set of data ***bivariate data***. Often when we have a set of bivariate data, we can control one of the quantities. We generally denote the quantity that can be controlled, the ***independent variable***, x. The other variable, the ***dependent variable***, is denoted as y. In this problem, we will assume the number of defective parts produced is affected by the number of workers absent. Therefore, we will call the number of workers absent x and the number of defective parts produced y. When we plot bivariate data, the independent variable is marked on the horizontal axis and the dependent variable is marked on the vertical axis. Therefore, for this example, number of workers absent is marked on the horizontal axis and number of defective parts is marked on the vertical axis. If we plot the six pieces of bivariate data in the Cartesian coordinate system, we get a scatter diagram, as shown in Fig. 13.38.

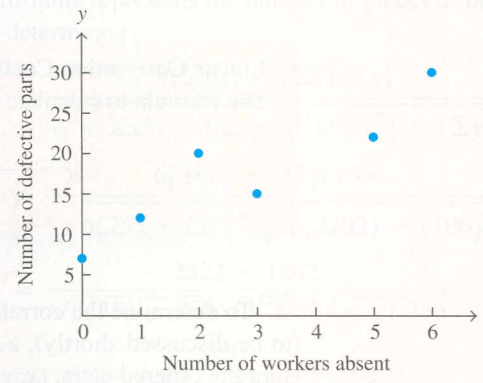


Figure 13.38

Figure 13.38 shows that, generally, the more workers that are absent, the more defective parts are produced.

In Fig. 13.39, we show some scatter diagrams and indicate the corresponding strength of correlation between the quantities on the horizontal and vertical axes.

Earlier we mentioned that r will always be a value between -1 and 1 inclusive. A value of $r = 1$ is obtained only when every point of the bivariate data on a scatter diagram lies in a straight line and the line is increasing from left to right (see Fig. 13.39a). In other words, the line has a positive slope, as discussed in Section 6.6.

A value of $r = -1$ will be obtained only when every point of the bivariate data on a scatter diagram lies in a straight line and the line is decreasing from left to right (see Fig. 13.39e). In other words, the line has a negative slope.

The value of r is a measure of how far a set of points varies from a straight line. The greater the spread, the weaker the correlation and the closer the value of r is to 0. Figure 13.39 shows that the more the dots diverge from a straight line, the weaker the correlation becomes.

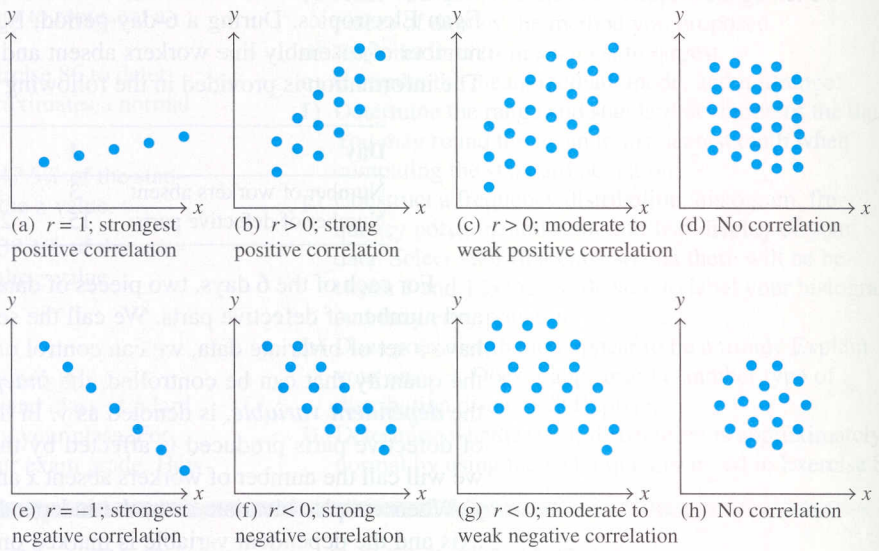


Figure 13.39

The following formula is used to calculate r.

Linear Correlation Coefficient

The formula to calculate the **correlation coefficient, r,** is as follows.

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{n(\Sigma x^2) - (\Sigma x)^2}\sqrt{n(\Sigma y^2) - (\Sigma y)^2}}$$

To determine the correlation coefficient, r, and the equation of the line of best fit (to be discussed shortly), a statistical calculator may be used. On the calculator you enter the ordered pairs, (x, y), and press the appropriate keys.

In Example 1, we show how to determine r for a set of bivariate data without the use of a statistical calculator. We will use the same set of bivariate data given on page 809 that was used to make the scatter diagram in Figure 13.38.

EXAMPLE 1 *Number of Absences versus Number of Defective Parts*

Egan Electronics provided the following daily records about the number of assembly line workers absent and the number of defective parts produced for 6 days. Determine the correlation coefficient between the number of workers absent and the number of defective parts produced.

Day	1	2	3	4	5	6
Number of workers absent	3	5	0	1	2	6
Number of defective parts	15	22	7	12	20	30

SOLUTION: We plotted this set of data on the scatter diagram in Figure 13.38. We will call the number of workers absent x. We will call the number of defective parts produced y. We list the values of x and y and calculate the necessary sums: Σx, Σy, Σxy, Σx^2, Σy^2. We determine the values in the column labeled x^2 by squaring the x's (multiplying the x's by themselves). We determine the values in the column labeled y^2 by squaring the y's. We determine the values in the column labeled xy by multiplying each x value by its corresponding y value.

Number of Workers Absent	Number of Defective Parts			
x	y	x^2	y^2	xy
3	15	9	225	45
5	22	25	484	110
0	7	0	49	0
1	12	1	144	12
2	20	4	400	40
6	30	36	900	180
17	106	75	2202	387

Thus, $\Sigma x = 17$, $\Sigma y = 106$, $\Sigma x^2 = 75$, $\Sigma y^2 = 2202$, and $\Sigma xy = 387$. In the formula for r, we use both $(\Sigma x)^2$ and Σx^2. Note that $(\Sigma x)^2 = (17)^2 = 289$ and that $\Sigma x^2 = 75$. Similarly, $(\Sigma y)^2 = (106)^2 = 11{,}236$ and $\Sigma y^2 = 2202$.

The n in the formula represents the number of pieces of bivariate data. Here $n = 6$. Now let's determine r.

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{n(\Sigma x^2) - (\Sigma x)^2}\sqrt{n(\Sigma y^2) - (\Sigma y)^2}}$$

$$= \frac{6(387) - (17)(106)}{\sqrt{6(75) - (17)^2}\sqrt{6(2202) - (106)^2}}$$

$$= \frac{2322 - 1802}{\sqrt{6(75) - 289}\sqrt{6(2202) - 11{,}236}}$$

$$= \frac{520}{\sqrt{450 - 289}\sqrt{13{,}212 - 11{,}236}}$$

$$= \frac{520}{\sqrt{161}\sqrt{1976}} \approx 0.92$$

TABLE 13.8 Correlation Coefficient, r

n	$\alpha = 0.05$	$\alpha = 0.01$
4	0.950	0.990
5	0.878	0.959
6	0.811	0.917
7	0.754	0.875
8	0.707	0.834
9	0.666	0.798
10	0.632	0.765
11	0.602	0.735
12	0.576	0.708
13	0.553	0.684
14	0.532	0.661
15	0.514	0.641
16	0.497	0.623
17	0.482	0.606
18	0.468	0.590
19	0.456	0.575
20	0.444	0.561
22	0.423	0.537
27	0.381	0.487
32	0.349	0.449
37	0.325	0.418
42	0.304	0.393
47	0.288	0.372
52	0.273	0.354
62	0.250	0.325
72	0.232	0.302
82	0.217	0.283
92	0.205	0.267
102	0.195	0.254

The derivation of this table is beyond the scope of this text. It shows the critical values of the Pearson correlation coefficient.

Since the maximum possible value for r is 1.00, a correlation coefficient of 0.92 is a strong, positive correlation. This result implies that, generally, the more assembly line workers absent, the more defective parts produced. ▲

In Example 1, had we found r to be a value greater than 1 or less than -1, it would have indicated that we had made an error. Also, from the scatter diagram, we should realize that r should be a positive value and not negative.

In Example 1, there appears to be a cause–effect relationship. That is, the more assembly line workers who are absent, the more defective parts are produced. *However, a correlation does not necessarily indicate a cause–effect relationship.* For example, there is a positive correlation between police officers' salaries and the cost of medical insurance over the past 10 years (both have increased), but that does not mean that the increase in police officers' salaries caused the increase in the cost of medical insurance.

Suppose in Example 1 that r had been 0.53. Would this value have indicated a correlation? What is the minimum value of r needed to assume that a correlation exists between the variables? To answer this question, we introduce the term *level of significance*. The ***level of significance***, denoted α (alpha), is used to identify the cutoff between results attributed to chance and results attributed to an actual relationship between the two variables. Table 13.8 gives ***critical values**** (or cutoff values) that are sometimes used for determining whether two variable are related. The table indicates two different levels of significance: $\alpha = 0.05$ and $\alpha = 0.01$. A level of significance of 5%, written $\alpha = 0.05$, means that there is a 5% chance that, when you say the variables are related, they actually are *not* related. Similarly, a level of significance of 1%, or $\alpha = 0.01$, means that there is a 1% chance that, when you say the variables are related, they actually are *not* related. More complete critical value tables are available in statistics books.

To explain the use of the table, we use ***absolute value***, symbolized $|\ |$. The absolute value of a nonzero number is the positive value of the number and the absolute value of 0 is 0. Therefore,

$$|3| = 3, \quad |-3| = 3, \quad |5| = 5, \quad |-5| = 5, \quad \text{and} \quad |0| = 0$$

If the absolute value of r, written $|r|$, is *greater than* the value given in the table under the specified α and appropriate sample size n, we assume that a correlation does exist between the variables. If $|r|$ is less than the table value, we assume that no correlation exists.

Returning to Example 1, if we want to determine whether there is a correlation at a 5% level of significance, we find the critical value (or cutoff value) that corresponds to $n = 6$ (there are 6 pieces of bivariate data) and $\alpha = 0.05$. The value to the right of $n = 6$ and under the $\alpha = 0.05$ column is the critical value 0.811. From the formula, we had obtained $r = 0.92$. Since $|0.92| > 0.811$, or $0.92 > 0.811$, we assume that a correlation between the variables exists.

Note in Table 13.8 that the larger the sample size, the smaller is the value of r needed for a significant correlation.

EXAMPLE 2 *Amount of Drug Remaining in the Bloodstream*

To test the length of time that an infection-fighting drug stays in a person's bloodstream, a doctor gives 300 milligrams of the drug to 10 patients labeled 1–10 in the table on page 813. Once each hour, for 8 hours, one of the 10 patients is selected at

*This table of values may be used only under certain conditions. If you take a statistics course, you will learn more about which critical values to use to determine whether a linear correlation exists.

random and that person's blood is tested to determine the amount of the drug remaining in the bloodstream. The results are as follows.

Patient	1	2	3	4	5	6	7	8	9	10
Time (hr)	1	2	3	4	5	6	7	8	9	10
Drug remaining (mg)	250	230	200	210	140	120	210	100	90	85

Determine at a level of significance of 5% whether a correlation exists between the time elapsed and the amount of drug remaining.

SOLUTION: Let time be represented by x and the amount of drug remaining by y. We first draw a scatter diagram (Fig. 13.40).

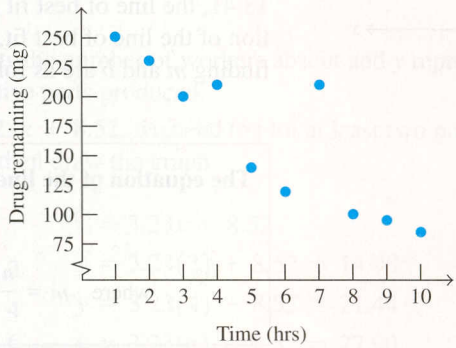


Figure 13.40

The scatter diagram suggests that, if a correlation exists, it will be negative. We now construct a table of values and calculate r.

x	y	x^2	y^2	xy
1	250	1	62,500	280
2	230	4	52,900	460
3	200	9	40,000	600
4	210	16	44,100	840
5	140	25	19,600	700
6	120	36	14,400	720
7	210	49	44,100	1470
8	100	64	10,000	800
9	90	81	8100	810
10	85	100	7225	850
55	1635	385	302,925	7500

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{n(\Sigma x^2) - (\Sigma x)^2}\sqrt{n(\Sigma y^2) - (\Sigma y)^2}}$$

$$= \frac{10(7500) - (55)(1635)}{\sqrt{10(385) - (55)^2}\sqrt{10(302{,}925) - (1635)^2}}$$

$$= \frac{-14{,}925}{\sqrt{825}\sqrt{355{,}025}} \approx \frac{-14{,}925}{17{,}114.19}$$

$$\approx -0.872$$

From Table 13.8, for $n = 10$ and $\alpha = 0.05$, we get 0.632. Since $|-0.872| = 0.872$ and $0.872 > 0.632$, a correlation exists. The correlation is negative, which indicates that the longer the time period, the smaller is the amount of drug remaining. ▲

Linear Regression

Let's now turn to regression. ***Linear regression*** is the process of determining the linear relationship between two variables. Recall from Section 6.6 that the slope–intercept form of a straight line is $y = mx + b$, where m is the slope and b is the y-intercept.

Using the set of bivariate data, we will determine the equation of ***the line of best fit.*** The line of best fit is also called ***the regression line,*** or ***the least squares line.*** The *line of best fit* is the line such that the sum of the vertical distances from the line to the data points (on the scatter diagram) is a minimum, as shown in Fig. 13.41. In Fig. 13.41, the line of best fit minimizes the sum of d_1 through d_8. To determine the equation of the line of best fit, $y = mx + b$ * we must find m and then b. The formulas for finding m and b are as follows.

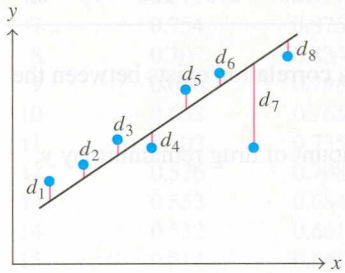


Figure 13.41

The equation of the line of best fit is

$$y = mx + b,$$

where $$m = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{n(\Sigma x^2) - (\Sigma x)^2}, \quad \text{and} \quad b = \frac{\Sigma y - m(\Sigma x)}{n}$$

Note that the numerator of the fraction used to find m is identical to the numerator used to find r. Therefore, if you have previously found r, you do not need to repeat this calculation. Also, the denominator of the fraction used to find m is identical to the radicand of the first square root in the denominator of the fraction used to find r.

EXAMPLE 3 *The Line of Best Fit*

a) Use the data in Example 1 to find the equation of the line of best fit that relates the number of workers absent on an assembly line and the number of defective parts produced.

b) Graph the equation of the line of best fit on a scatter diagram that illustrates the set of bivariate points.

SOLUTION:

a) In Example 1, we found $n(\Sigma xy) - (\Sigma x)(\Sigma y) = 520$ and $n(\Sigma x^2) - (\Sigma x)^2 = 161$. Thus,

$$m = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{n(\Sigma x^2) - (\Sigma x)^2} = \frac{520}{161} \approx 3.23$$

*Some statistics books use $y = ax + b$, $y = b_0 + b_1x$, or something similar for the equation of the line of best fit. In any case, the letter next to the variable x represents the slope of the line of best fit, and the other letter represents the y-intercept of the graph.

Now we find the y-intercept, b. In Example 1, we found $n = 6$, $\Sigma x = 17$, and $\Sigma y = 106$.

$$b = \frac{\Sigma y - m(\Sigma x)}{n}$$
$$\approx \frac{106 - 3.23(17)}{6} \approx \frac{51.09}{6} \approx 8.52$$

Therefore, the equation of the line of best fit is

$$y = mx + b$$
$$y = 3.23x + 8.52$$

where x represents the number of workers absent and y represents the predicted number of defective parts produced.

b) To graph $y = 3.23x + 8.52$, we need to plot at least two points. We will plot three points and then draw the graph.

$y = 3.23x + 8.52$

		x	y
$x = 2$	$y = 3.23(2) + 8.52 = 14.98$	2	14.98
$x = 4$	$y = 3.23(4) + 8.52 = 21.44$	4	21.44
$x = 6$	$y = 3.23(6) + 8.52 = 27.90$	6	27.90

These three calculations indicate that if 2 assembly line workers are absent on the assembly line, the predicted number of defective parts produced is about 15. If 4 assembly line workers are absent, the predicted number of defective parts produced is about 21, and if 6 assembly line workers are absent, the predicted number of defective parts produced is about 28. Plot the three points (the three red points in Figure 13.42) and then draw a straight line through the three points. The scatter diagram and graph of the equation of the line of best fit are plotted in Fig. 13.42.

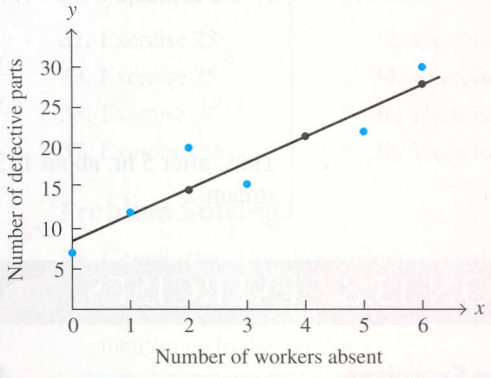


Figure 13.42

In Example 3, the line of best fit intersects the y-axis at 8.52, the value we determined for b in part (a).

EXAMPLE 4 *Line of Best Fit for Example 2*

a) Determine the equation of the line of best fit between the time elapsed and the amount of drug remaining in a person's bloodstream in Example 2 on pages 812 and 813.

b) If the average person is given 300 mg of the drug, how much will remain in the person's bloodstream after 5 hr?

SOLUTION:

a) From the scatter diagram on page 813 we see that the slope of the line of best fit, m, will be negative. In Example 2, we found that $n(\Sigma xy) - (\Sigma x)(\Sigma y) = -14{,}925$ and that $n(\Sigma x^2) - (\Sigma x)^2 = 825$. Thus,

$$m = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{n(\Sigma x^2) - (\Sigma x)^2}$$
$$= \frac{-14{,}925}{825}$$
$$\approx -18.1$$

From Example 2, $n = 10$, $\Sigma x = 55$, and $\Sigma y = 1635$.

$$b = \frac{\Sigma y - m\Sigma x}{n}$$
$$= \frac{1635 - (-18.1)(55)}{10}$$
$$\approx 263.1$$

Thus, the equation of the line of best fit is

$$y = mx + b$$
$$y = -18.1x + 263.1$$

where x is the elapsed time and y is the amount of drug remaining.

b) We evaluate $y = -18.1x + 263.1$ at $x = 5$.

$$y = -18.1x + 263.1$$
$$y = -18.1(5) + 263.1 = 172.6$$

Thus, after 5 hr, about 173 mg of the drug remains in the average person's bloodstream. ▲

SECTION 13.8 EXERCISES

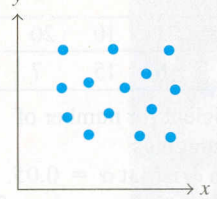
Concept/Writing Exercises

1. What does the correlation coefficient measure?
2. What is the purpose of regression?
3. What value of r represents the maximum positive correlation?
4. What value of r represents the maximum negative correlation?
5. What value of r represents no correlation between the variables?
6. What does a negative correlation between two variables indicate?

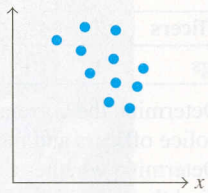
7. What does a positive correlation between two variables indicate?
8. What does the line of best fit represent?
9. What does the level of significance signify?
10. What is a scatter diagram?

In Exercises 11–14, indicate if you believe that a correlation exists between the quantities on the horizontal and vertical axis. If so, indicate if you believe that the correlation is a strong positive correlation, a strong negative correlation, a weak positive correlation, a weak negative correlation, or no correlation. Explain your answer.

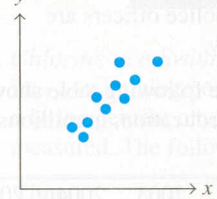
11.



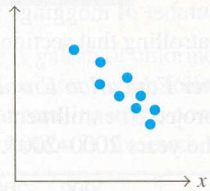
12.



13.



14.



Practice the Skills

In Exercises 15–22, assume that a sample of bivariate data yields the correlation coefficient, r, indicated. Use Table 13.8 on page 812 for the specified sample size and level of significance to determine whether a linear correlation exists.

15. $r = 0.76$ when $n = 13$ at $\alpha = 0.01$
16. $r = 0.43$ when $n = 22$ at $\alpha = 0.01$
17. $r = -0.73$ when $n = 8$ at $\alpha = 0.05$
18. $r = -0.49$ when $n = 11$ at $\alpha = 0.05$
19. $r = -0.23$ when $n = 102$ at $\alpha = 0.01$
20. $r = -0.49$ when $n = 18$ at $\alpha = 0.01$
21. $r = 0.82$ when $n = 6$ at $\alpha = 0.01$
22. $r = 0.96$ when $n = 5$ at $\alpha = 0.01$

In Exercises 23–30, (a) draw a scatter diagram; (b) find the value of r, rounded to the nearest thousandth; (c) determine whether a correlation exists at $\alpha = 0.05$; and (d) determine whether a correlation exists at $\alpha = 0.01$.

23.

x	y
4	6
5	9
6	11
7	11
10	13

24.

x	y
6	12
8	11
11	9
14	10
17	8

25.

x	y
23	29
35	37
31	26
43	20
49	39

26.

x	y
90	3
80	4
60	6
60	5
40	5
20	7

27.

x	y
5.3	10.3
4.7	9.6
8.4	12.5
12.7	16.2
4.9	9.8

28.

x	y
12	15
16	19
13	45
24	30
100	60
50	28

29.

x	y
100	2
80	3
60	5
60	6
40	6
20	8

30.

x	y
90	90
70	70
65	65
60	60
50	50
40	40
15	15

In Exercises 31–38, determine the equation of the line of best fit from the data in the exercise indicated. Round both the slope and the y-intercept to the nearest tenth.

31. Exercise 23
32. Exercise 24
33. Exercise 25
34. Exercise 26
35. Exercise 27
36. Exercise 28
37. Exercise 29
38. Exercise 30

Problem Solving

39. *Commuting Time* Six students provided the following data about the distance (in miles) from their home to their college and the time (in minutes) required to commute from their home to their college.

Distance (miles)	8	20	9	15	16	2
Time (minutes)	15	28	20	25	28	5

a) Determine the correlation coefficient between the distance and the time to commute.
b) Determine whether a correlation exists at $\alpha = 0.05$.

c) Find the equation of the line of best fit for the distance and the time to commute.

40. ***Amount of Fat in Pizza*** The January 2002 issue of *Consumer Reports* reported the following information regarding the number of calories and the number of grams of fat in a 5 oz slice of pizza for the top-rated pizza chains.

Calories	321	380	350	358	378	391
Fat (grams)	13	23	16	14	19	19

a) Determine the correlation coefficient between the number of calories and the number of grams of fat.
b) Determine whether a correlation exists at $\alpha = 0.05$.
c) Find the equation of the line of best fit for the number of calories and the number of grams of fat.

41. ***Amount of Time Spent Studying*** Six students provided the following data about the lengths of time they studied for a psychology exam and the grades they received on the exam.

Time studied (minutes)	20	40	50	60	80	100
Grade received (percent)	40	45	70	76	92	95

a) Determine the correlation coefficient between the length of time studied and the grade received.
b) Determine whether a correlation exists at $\alpha = 0.01$.
c) Find the equation of the line of best fit for the length of time studied and the grade received.

42. ***Hiking*** The following table shows the number of hiking permits issued for a specific trail at Yellowstone National Park for selected years and the corresponding number of mountain lions sighted by the hikers on that trail.

Hiking permits	765	926	1145	842	1485	1702
Mountain lions	119	127	150	119	153	156

a) Determine the correlation coefficient between the number of hiking permits issued and the number of mountain lions sighted by hunters.
b) Determine whether a correlation exists at $\alpha = 0.05$.
c) Determine the equation of the line of best fit for the number of hiking permits issued and the number of mountain lions sighted by hunters.
d) Use the equation in part (c) to estimate the number of mountain lions sighted by hunters if 1500 hiking permits were issued.

43. ***City Muggings*** In a certain section of a city, muggings have been a problem. The number of police officers patrolling that section of the city has varied. The following chart shows the number of police officers and the number of muggings for 10 successive days.

Police officers	20	12	18	15	22	10	20	12
Muggings	8	10	12	9	6	15	7	18

a) Determine the correlation coefficient for number of police officers and number of muggings.
b) Determine whether a correlation exists at $\alpha = 0.05$.
c) Find the equation of the line of best fit for number of police officers and number of muggings.
d) Use the equation in part (c) to estimate the average number of muggings when 14 police officers are patrolling that section of the city.

44. ***Higher Education Enrollment*** The following table shows the projected enrollment in higher education, in millions, for the years 2000–2005.

Year	2000	2001	2002	2003	2004	2005
Enrollment (in millions)	15.0	15.3	15.5	15.8	16.1	16.3

Source: U.S. Center for Educational Statistics

a) Determine the correlation coefficient between the year and the projected enrollment in higher education.
b) Determine whether a correlation exists at $\alpha = 0.05$.
c) Find the equation of the line of best fit for the year and the projected enrollment in higher education.
d) Use the equation in part (c) to estimate the projected enrollment in higher education in 2008.

45. ***Selling Popcorn at the Movies*** The number of movie tickets sold and the number of units of popcorn sold at Regal Cinema for 8 days is shown below.

Ticket sales	89	110	125	92	100	95	108	97
Units of popcorn	22	28	30	26	22	21	28	25

a) Determine the correlation coefficient between ticket sales and units of popcorn sold.
b) Determine whether a correlation exists at $\alpha = 0.05$.
c) Determine the equation of the line of best fit for tickets sold and units of popcorn sold.
d) Use the equation in part (c) to estimate the units of popcorn sold if 115 tickets are sold.

46. ***Movie Ratings*** A popular newspaper rates movies from one to four stars (four stars is the highest rating). The following table shows the ratings of 10 movies selected at random and the gross earnings of each movie.

Rating (stars)	4	4	3	2	1	3	4	2	4	1
Earnings (millions of dollars)	100	67	80	120	40	90	60	60	90	100

a) Determine the correlation coefficient between number of stars and the movies' earnings.
b) Determine whether a correlation exists at $\alpha = 0.05$.
c) Determine the equation of the line of best fit for the number of stars and the movies' earnings.

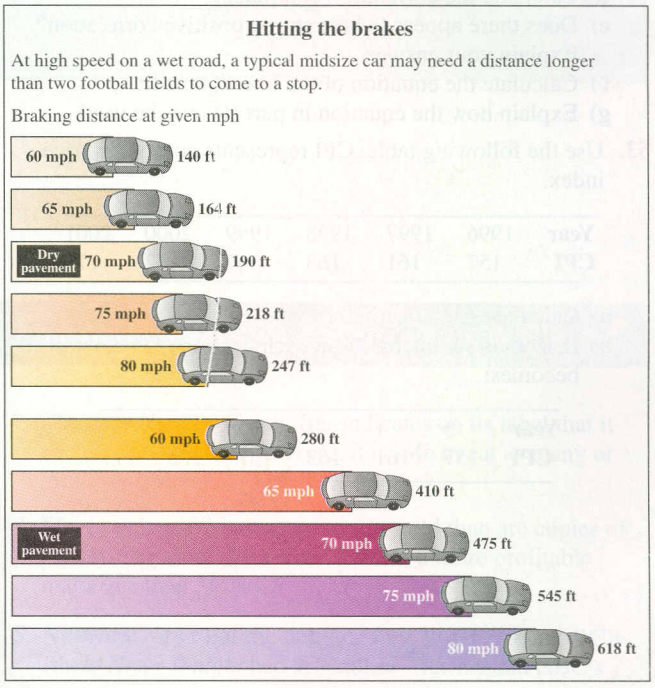
47. ***Chlorine in a Swimming Pool*** A gallon of chlorine is put into a swimming pool. Each hour later for the following 6 hr the percent of chlorine that remains in the pool is measured. The following information is obtained.

Time	1	2	3	4	5	6
Chlorine remaining (percent)	80.0	76.2	68.7	50.1	30.2	20.8

a) Determine the correlation coefficient for time and percent of chlorine remaining.
b) Determine whether a correlation exists at $\alpha = 0.01$.
c) Determine the equation of the line of best fit for time and amount of chlorine remaining.
d) Use the equation in part (c) to estimate the average amount of chlorine remaining after 4.5 hr.

48. ***Social Security Numbers*** **a)** Match the first 9 digits in your phone number (including area code) with the 9 digits in your social security number. To do so, match the first digit in your phone number with the first digit in your social security number to get one ordered pair. Match the second digits to get a second ordered pair. Continue this process until you get a total of nine ordered pairs.
b) Do you believe that this set of bivariate data has a positive correlation, a negative correlation, or no correlation? Explain your answer.
c) Construct a scatter diagram for the nine ordered pairs.
d) Calculate the correlation coefficient, r.
e) Is there a correlation at $\alpha = 0.05$? Explain.
f) Calculate the equation of the line of best fit.
g) Use the equation in part (f) to estimate the digit in a social security number that corresponds with a 7 in a telephone number.

49. ***Hitting the Brakes*** **a)** Examine the art below. Do you believe that there is a positive correlation, a negative correlation, or no correlation between speed of a car and stopping distance when the brakes are applied? Explain.
b) Do you believe that there is a stronger correlation between speed of a car and stopping distance on wet or dry roads? Explain.
c) Use the figure to construct two scatter diagrams, one for dry pavement and the other for wet pavement. Place the speed of the car on the horizontal axis.
d) Compute the correlation coefficient for speed of the car and stopping distance for dry pavement.
e) Repeat part (d) for wet pavement.
f) Were your answers to parts (a) and (b) correct? Explain.
g) Determine the equation of the line of best fit for dry pavement.
h) Repeat part (g) for wet pavement.



Source: Car and Driver, American Automobile Association

i) Use the equations in parts (g) and (h) to estimate the stopping distance of a car going 77 mph on both dry and wet pavements.

Challenge Problems/Group Activities

50. a) Assume that a set of bivariate data yields a specific correlation coefficient. If the x and y values are interchanged and the correlation coefficient is recalculated, will the correlation coefficient change? Explain.

b) Make up a table of five pieces of bivariate data and determine r using the data. Then switch the values of the x's and y's and recompute the correlation coefficient. Has the value of r changed?

51. a) Do you believe that a correlation exists between a person's height and the length of a person's forearm? Explain.

b) Select 10 people from your class and measure (in inches) their heights and the lengths of their forearms.

c) Plot the 10 ordered pairs on a scatter diagram.

d) Calculate the correlation coefficient, r.

e) Determine the equation of the line of best fit.

f) Estimate the length of the forearm of a person who is 58 in. tall.

52. a) Have your group select a category of bivariate data that it thinks has a strong positive correlation. Designate the independent variable and the dependent variable. Explain why your group believes that the bivariate data have a strong positive correlation.

b) Collect at least 10 pieces of bivariate data that can be used to determine the correlation coefficient. Explain how your group chose these data.

c) Plot a scatter diagram.

d) Calculate the correlation coefficient.

e) Does there appear to be a strong positive correlation? Explain your answer.

f) Calculate the equation of the line of best fit.

g) Explain how the equation in part (f) may be used.

53. Use the following table. CPI represents consumer price index.

Year	1996	1997	1998	1999	2000	2001
CPI	157	161	163	167	172	177

a) Calculate r.

b) If 1996 is subtracted from each year, the table obtained becomes:

Year	0	1	2	3	4	5
CPI	157	161	163	167	172	177

If r is calculated from these values, how will it compare with the r determined in part (a)? Explain.

c) Calculate r from the values in part (b) and compare the results with the value of r found in part (a). Are they the same? If not, explain why.

54. a) There are equivalent formulas that can be used to find the correlation coefficient and the equation of the line of best fit. A formula used in some statistics books to find the correlation coefficient is

$$r = \frac{SS(xy)}{\sqrt{SS(x)SS(y)}}$$

where

$$SS(x) = \Sigma x^2 - \frac{(\Sigma x)^2}{n}$$

$$SS(y) = \Sigma y^2 - \frac{(\Sigma y)^2}{n}$$

$$\text{and} \quad SS(xy) = \Sigma xy - \frac{(\Sigma x)(\Sigma y)}{n}$$

Use this formula to find the correlation coefficient of the set of bivariate data given in Example 1 on page 811.

b) Compare your answer with the answer obtained in Example 1.

Internet/Research Activities

55. a) Obtain a set of bivariate data from a newspaper or magazine.

b) Plot the information on a scatter diagram.

c) Indicate whether you believe that the data show a positive correlation, a negative correlation, or no correlation. Explain your answer.

d) Calculate r and determine whether your answer to part (c) was correct.

e) Determine the equation of the line of best fit for the bivariate data.

56. Find a scatter diagram in a newspaper or magazine and write a paper on what the diagram indicates. Indicate whether you believe that the bivariate data show a positive correlation, a negative correlation, or no correlation and explain why.

CHAPTER 13 SUMMARY

IMPORTANT FACTS

Rules for data grouped by classes

1. The classes should be the same width.
2. The classes should not overlap.
3. Each piece of data should belong to only one class.

Measures of central tendency

The **mean** is the sum of the data divided by the number of pieces of data: $\bar{x} = \dfrac{\Sigma x}{n}$.

The **median** is the value in the middle of a set of ranked data.

The **mode** is the piece of data that occurs most frequently (if there is one).

The **midrange** is the value halfway between the lowest and highest values: $\text{midrange} = \dfrac{L + H}{2}$.

Statistical graphs

Circle graph

Histogram

Frequency polygon

Stem-and-leaf display

Measures of dispersion

The **range** is the difference between the highest value and lowest value in a set of data.

The **standard deviation,** s, is a measure of the spread of a set of data about the mean: $s = \sqrt{\dfrac{\Sigma(x - \bar{x})^2}{n - 1}}$.

z-scores

$$z = \frac{x - \mu}{\sigma}$$

Chebyshev's theorem

$$\text{Minimum percentage} = 1 - \frac{1}{K^2},\ K > 1$$

Linear correlation and regression

Linear correlation coefficient, r, is

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{n(\Sigma x^2) - (\Sigma x)^2}\sqrt{n(\Sigma y^2) - (\Sigma y)^2}}$$

Equation of the line of the best fit

$y = mx + b$, where

$$m = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{n(\Sigma x^2) - (\Sigma x)^2}, \text{ and}$$

$$b = \frac{\Sigma y - m(\Sigma x)}{n}$$

CHAPTER 13 REVIEW EXERCISES

13.1

1. **a)** What is a population?
   b) What is a sample?
2. What is a random sample?

13.2

In Exercises 3 and 4, tell what possible misuses or misinterpretations may exist in the statements.

3. The Stay Healthy Candy Bar indicates on its label that it has no cholesterol. Therefore, it is safe to eat as many of these candy bars as you want.
4. More copies of *Time* magazine are sold than are copies of *Money* magazine. Therefore, *Time* is a more profitable magazine than *Money*.
5. ***Network News*** For the week of July 15–21, 2002, ABC's *World News Tonight* had 9.1 million viewers and NBC's

Nightly News had 8.9 million viewers. Draw a graph that appears to show

a) small difference in the number of viewers.

b) a large difference in the number of viewers.

13.3, 13.4

6. Consider the following set of data.

35	37	38	41	43
36	37	38	41	43
36	37	39	41	43
36	37	39	41	44
37	37	39	42	45

a) Construct a frequency distribution letting each class have a width of 1.

b) Construct a histogram.

c) Construct a frequency polygon.

7. ***Average Daily High Temperature*** Consider the following average daily high temperature in January for 40 selected cities.

87	41	54	34	89	83	65	80
69	75	47	43	68	67	48	35
86	30	56	44	50	33	77	77
42	84	75	54	81	86	79	66
48	42	73	83	36	88	66	55

a) Construct a frequency distribution. Let the first class be 30–39.

b) Construct a histogram of the frequency distribution.

c) Construct a frequency polygon of the frequency distribution.

d) Construct a stem-and-leaf display.

13.5, 13.6

In Exercises 8–13, for the following test scores 63, 76, 79, 83, 86, 93, determine the

8. mean. **9.** median.

10. mode. **11.** midrange.

12. range. **13.** standard deviation.

In Exercises 14–19, for the set of data 4, 5, 12, 14, 19, 7, 12, 23, 7, 17, 15, 21, determine the

14. mean. **15.** median.

16. mode. **17.** midrange.

18. range. **19.** standard deviation.

13.7

Anthropology *In Exercises 20–24, assume that anthropologists have determined that a certain type of primitive animal had a mean head circumference of 42 cm with a standard deviation of 5 cm. Given that head sizes were normally distributed, determine the percent of heads*

20. between 37 and 47 cm. **21.** between 32 and 52 cm.

22. less than 50 cm. **23.** greater than 50 cm.

24. greater than 39 cm.

Pizza Delivery *In Exercises 25–28, assume that the amount of time to prepare and deliver a pizza from Dominos is normally distributed with a mean of 20 min and standard deviation of 5 min. Find the percent of pizzas that were prepared and delivered*

25. between 20 and 25 min.

26. in less than 18 min.

27. between 22 and 28 min.

28. If Dominos advertises that the pizza is free if it takes more than 30 min to deliver, what percent of the pizza will be free?

13.8

In Exercises 29 and 30, use the following table that shows both the U.S. sales of flowers and plants and the number of greenhouse owners in New York State, where the column labeled as Year refers to the number of years since 1997. Thus 1997 would correspond to year 0.

Year	U.S. Flower and Plant Sales (billions)	Greenhouse Owners in New York State
0	$3.9	897
1	3.9	800
2	4.1	770
3	4.6	760
4	4.8	735
5	4.9	663

29. a) ***Flower and Plant Sales*** Construct a scatter diagram for the U.S. sales of flowers and plants with the years, from 0 through 5, on the horizontal axis.

b) Use the scatter diagram in part (a) to determine whether you believe a correlation exists between the year and U.S. sales of flowers and plants. If so, is it a positive or a negative correlation? Explain.

c) Calculate the correlation coefficient between the year and the U.S. sales of flowers and plants.

d) Determine whether a correlation exists at $\alpha = 0.05$.

e) Determine the equation of the line of best fit between the year and U.S. sales of flowers and plants.

30. ***Greenhouse Owners in New York State***

a) Construct a scatter diagram for the number of greenhouse owners in New York State with the years, from 0 through 5, on the horizontal axis.

b) Use the scatter diagram in part (a) to determine whether you believe a correlation exists between the year and the number of greenhouse owners in New York State. If so, is it a positive or a negative correlation? Explain.
c) Calculate the correlation coefficient between the year and the number of greenhouse owners in New York State.
d) Determine whether a correlation exists at $\alpha = 0.05$.
e) Determine the equation of the line of best fit between the year and the number of greenhouse owners in New York State.

31. ***Daily Sales*** Ace Hardware recorded the number of a particular item sold per week for 6 weeks and the corresponding weekly price, in dollars, of the item as shown in the table below.

Price ($)	0.75	1.00	1.25	1.50	1.75	2.00
Number sold	200	160	140	120	110	95

a) Construct a scatter diagram with price on the horizontal axis.
b) Use the scatter diagram in part (a) to determine whether you believe that a correlation exists between the price of the item and number sold. If so, it is a positive or a negative correlation? Explain.
c) Determine the correlation coefficient between the price and the number sold.
d) Determine whether a correlation exists at $\alpha = 0.05$. Explain how you arrived at your answer.
e) Determine the equation of the line of best fit for the price and the number sold.
f) Use the equation in part (e) to estimate the number sold if the price is $1.60.

13.5–13.7

Men's Weight *In Exercises 32–39, use the following data obtained from a study of the weights of adult men.*

Mean	187 lb	First quartile	173 lb
Median	180 lb	Third quartile	227 lb
Mode	175 lb	86th percentile	234 lb
Standard deviation	23 lb		

32. What is the most common weight?
33. What weight did half of those surveyed exceed?
34. About what percent of those surveyed weighed more than 227 lb?
35. About what percent of those surveyed weighed less than 173 lb?
36. About what percent of those surveyed weighed more than 234 lb?
37. If 100 men were surveyed, what is the total weight of all men?
38. What weight represents two standard deviations above the mean?
39. What weight represents 1.8 standard deviations below the mean?

13.2–13.7

Presidential Children *The following list shows the names of the 42 U.S. presidents and the number of children in their families.*

Washington	0	Cleveland	5
J. Adams	5	B. Harrison	3
Jefferson	6	McKinley	2
Madison	0	T. Roosevelt	6
Monroe	2	Taft	3
J. Q. Adams	4	Wilson	3
Jackson	0	Harding	0
Van Buren	4	Coolidge	2
W. H. Harrison	10	Hoover	2
Tyler	14	F. D. Roosevelt	6
Polk	0	Truman	1
Taylor	6	Eisenhower	2
Fillmore	2	Kennedy	3
Pierce	3	L. B. Johnson	2
Buchanan	0	Nixon	2
Lincoln	4	Ford	4
A. Johnson	5	Carter	4
Grant	4	Reagan	4
Hayes	8	G. Bush	6
Garfield	7	Clinton	1
Arthur	3	G. W. Bush	2

In Exercises 40–51, use the data above to determine the following.

40. Mean number of children
41. Mode
42. Median
43. Midrange
44. Range
45. Standard deviation (round the mean to the nearest tenth)
46. Construct a frequency distribution; let the first class be 0–1.
47. Construct a histogram.
48. Construct a frequency polygon.
49. Does this distribution appear to be normal? Explain.
50. On the basis of this sample, do you think the number of children per family in the United States is a normal distribution? Explain.
51. Do you believe that this sample is representative of the population? Explain.

CHAPTER 13 TEST

In Exercises 1–6, for the set of data 21, 37, 37, 39, 46, determine the

1. mean.

2. median.

3. mode.

4. midrange.

5. range.

6. standard deviation.

In Exercises 7–9, use the set of data

26	28	35	46	49	56
26	30	36	46	49	58
26	32	40	47	50	58
26	32	44	47	52	62
27	35	46	47	54	66

to construct

7. a frequency distribution; let the first class be 25–30.

8. a histogram of the frequency distribution.

9. a frequency polygon of the frequency distribution.

Statistics on Salaries *In Exercises 10–16, use the following data on weekly salaries at Maxwell Mechanical Contractors.*

Mean	$700	First quartile	$650
Median	$670	Third quartile	$705
Mode	$695	79th percentile	$712
Standard deviation	$40		

10. What is the most common salary?

11. What salary did half the employees exceed?

12. About what percent of employees' salaries exceeded $650?

13. About what percent of employees' salaries was less than $712?

14. If the company has 100 employees, what is the total weekly salary of all employees?

15. What salary represents one standard deviation above the mean?

16. What salary represents 1.5 standard deviations below the mean?

Mileage of 5-Year-Old Cars *In Exercises 17–20, the mileage of 5-year-old cars is normally distributed with a mean of 75,000 and a standard deviation of 12,000 miles.*

17. What percent of 5-year-old cars have mileage between 50,000 and 70,000 miles?

18. What percent of 5-year-old cars have mileage greater than 60,000 miles?

19. What percent of 5-year-old cars have mileage greater than 90,000 miles?

20. If a random sample of 300 five-year-old cars is selected, how many would have mileage between 60,000 and 70,000 miles?

21. ***The Elderly U.S. Population*** The following chart shows the percent of the U.S. population that was age 65 and over for the years 1970, 1980, 1990, 1995, and 2000, where the column labeled Year refers to the number of years since 1970.

Percent of U.S. Population Age 65 and Over

Year	Percent
0	9.8
10	11.3
20	12.5
25	12.8
30	12.4

Source: U.S. Bureau of the Census, U.S. Dept. of Commerce.

a) Construct a scatter diagram placing the year on the horizontal axis.

b) Use the scatter diagram in part (a) to determine whether you believe that a correlation exists between the year and the percent of the U.S. population age 65 and over. Explain.

c) Determine the correlation coefficient between the year and the percent of the U.S. population age 65 and over.

d) Determine whether a correlation exists at $\alpha = 0.05$.

e) Assuming that this trend continues, determine the equation of the line of best fit between the year and the percent of the U.S. population age 65 and over.

f) Use the equation in part (e) to predict the percent of the U.S. population age 65 and over in 2010, or 40 years after 1970.

GROUP PROJECTS

Watching TV

1. Do you think that men or women, aged 17–20, watch more hours of TV weekly, or do you think that they watch the same number of hours?
   - **a)** Write a procedure to use to determine the answer to that question. In your procedure, use a sample of 30 men and 30 women. State how you will obtain an unbiased sample.
   - **b)** Collect 30 pieces of data from men aged 17–20 and 30 pieces of data from women aged 17–20. Round answers to the nearest 0.5 hr. Follow the procedure developed in part (a) to obtain your unbiased sample.
   - **c)** Compute the mean for your two groups of data to the nearest tenth.
   - **d)** Using the means obtained in part (c), answer the question asked at the beginning of the problem.
   - **e)** Is it possible that your conclusion in part (d) is wrong? Explain.
   - **f)** Compute the standard deviation for each group to the nearest tenth. How do the standard deviations compare?
   - **g)** Do you believe that the distribution of data from either or both groups resembles a normal distribution? Explain.
   - **h)** Add the two groups of data to get one group of 60 pieces of data. If these 60 pieces of data are added and divided by 60, will you obtain the same mean as when you add the two means from part (c) and divide the sum by 2? Explain.
   - **i)** Compute the mean of the 60 pieces of data by using both methods mentioned in part (h). Are they the same? If so, why? If not, why not?
   - **j)** Do you believe that this group of 60 pieces of data represents a normal distribution? Explain.

Binomial Probability Experiment

2. **a)** Have your group select a category of bivariate data that it thinks has a strong negative correlation. Indicate the variable that you will designate as the independent variable and the variable that you will designate as the dependent variable. Explain why your group believes that the bivariate data have a strong negative correlation.
   - **b)** Collect at least 10 pieces of bivariate data that can be used to determine the correlation coefficient. Explain how your group chose these data.
   - **c)** Plot a scatter diagram.
   - **d)** Calculate the correlation coefficient.
   - **e)** Is there a negative correlation at $\alpha = 0.05$? Explain your answer.
   - **f)** Calculate the equation of the line of best fit.
   - **g)** Explain how the equation in part (f) may be used.