

STATISTICS

Every year the admission officers of colleges choose, from thousands of applicants, those students who will be offered a place in the incoming class for the next year. An attempt is made by the college to choose students who will be best able to succeed academically and who best fit the profile of the student body of that college. Although this choice is based not only on academic standing, the scores on standardized tests are an important part of the selection. Statistics establishes the validity of the information obtained from standardized tests and influence the interpretation of the data obtained from them.

CHAPTER TABLE OF CONTENTS

- 15-1** Gathering Data
- 15-2** Measures of Central Tendency
- 15-3** Measures of Central Tendency for Grouped Data
- 15-4** Measures of Dispersion
- 15-5** Variance and Standard Deviation
- 15-6** Normal Distribution
- 15-7** Bivariate Statistics
- 15-8** Correlation Coefficient
- 15-9** Non-Linear Regression
- 15-10** Interpolation and Extrapolation
- Chapter Summary
- Vocabulary
- Review Exercises
- Cumulative Review

I 5-1 GATHERING DATA

Important choices in our lives are often made by evaluating information, but in order to use information wisely, it is necessary to organize and condense the multitude of facts and figures that can be collected. **Statistics** is the science that deals with the collection, organization, summarization, and interpretation of related information called **data**. **Univariate statistics** consists of one number for each data value.

Collection of Data

Where do data come from? Individuals, government organizations, businesses, and other political, scientific, and social groups usually keep records of their activities. These records provide factual data. In addition to factual data, the outcome of an event such as an election, the sale of a product, or the success of a movie often depends on the opinion or choices of the public.

Common methods of collecting data include the following:

1. **Censuses:** every ten years, the government conducts a *census* to determine the U.S. population. Each year, almanacs are published that summarize and update data of general interest.
2. **Surveys:** written questionnaires, personal interviews, or telephone requests for information can be used when experience, preference, or opinions are sought.
3. **Controlled experiments:** a structured study that usually consists of two groups: one that makes use of the subject of the study (for example, a new medicine) and a control group that does not. Comparison of results for the two groups is used to indicate effectiveness.
4. **Observational studies:** similar to controlled experiments except that the researcher does not apply the treatment to the subjects. For example, to determine if a new drug causes cancer, it would be unethical to give the drug to patients. A researcher *observes* the occurrence of cancer among groups of people who *previously* took the drug.

When information is gathered, it may include data for all cases to which the result of the study is to be applied. This source of information is called the **population**. When the *entire* population can be examined, the study is a **census**. For example, a study on the age and number of accidents of every driver insured by an auto insurance company would constitute a census if all of the company's records are included. However, when it is not possible to obtain information from every case, a **sample** is used to determine data that may then be applied to every case. For example, in order to determine the quality of their product, the quality-control department of a business may study a sample of the product being produced.

In order that the sample reflect the properties of the whole group, the following conditions should exist:

1. The sample must be representative of the group being studied.
2. The sample must be large enough to be effective.
3. The selection should be random or determined in such a way as to eliminate any bias.

If a new medicine being tested is proposed for use by people of all ages, of different ethnic backgrounds, and for use by both men and women, then the sample must be made up of people who represent these differences in sufficient number to be effective. A political survey to be effective must include people of different cultural, ethnic, financial, geographic, and political backgrounds.

Potential Pitfalls of Surveys

Surveys are a very common way of collecting data. However, if not done correctly, the results of the survey can be invalid. One potential problem is with the wording of the survey questions. For example, the question, “Do you agree that teachers should make more money?” will likely lead to a person answering “Yes.” A more neutral form of this question would be, “Do you believe that teachers’ salaries are too high, too low, or just about right?”

Questions can also be too vague, “loaded” (that is, use words with unintended connotations), or confusing. For example, for many people, words that invoke race will likely lead to an emotional response.

Another potential problem with surveys is the way that participants are selected. For example, a magazine would like to examine the typical teenager’s opinion on a pop singer in a given city. The magazine editors conduct a survey by going to a local mall. The problem with this survey is that teenagers who go to the mall are not necessarily representative of all teenagers in the city. A better survey would be done by visiting the high schools of the city.

Many surveys often rely on volunteers. However, volunteers are likely to have stronger opinions than the general population. This is why the selection of participants, if possible, should be random.

EXAMPLE I

A new medicine intended for use by adults is being tested on five men whose ages are 22, 24, 25, 27, and 30. Does the sample provide a valid test?

Solution No:

- The sample is too small.
- The sample includes only men.
- The sample does not include adults over 30.



EXAMPLE 2

The management of a health club has received complaints about the temperature of the water in the swimming pool. They want to sample 50 of the 200 members of the club to determine if the temperature of the pool should be changed. How should this sample be chosen?

- Solution** One suggestion might be to poll the first 50 people who use the pool on a given day. However, this will only include people who use the pool and who can therefore tolerate the water temperature.
- Another suggestion might be to place 50 questionnaires at the entrance desk and request members to respond. However, this includes only people who choose to respond and who therefore may be more interested in a change.
- A third suggestion might be to contact by phone every fourth person on the membership list and ask for a response. This method will produce a random sample but will include people who have no interest in using the pool. This sample may be improved by eliminating the responses of those people.

Organization of Data

In order to be more efficiently presented and more easily understood and interpreted, the data collected must be organized and summarized. Charts and graphs such as the histogram are useful tools. The **stem-and-leaf diagram** is an effective way of organizing small sets of data.

For example, the heights of the 20 children in a seventh-grade class are shown to the right. To draw a stem-and-leaf diagram, choose the tens digit as the *stem* and the units digit as the *leaf*.

Heights of Children									
61	71	58	72	60	53	74	61	68	65
72	67	64	48	70	56	65	67	59	61

- (1) Draw a vertical line and list the tens digits, 4, 5, 6, and 7 (or the **stem**), from bottom to top to the left of the line:
- (2) Enter each height by writing the **leaf**, the units digit, to the right of the line, following the appropriate stem:

Stem	
7	
6	
5	
4	

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

- (3) Arrange the leaves in numerical order after each stem:
- (4) Add a key to indicate the meaning of the numbers in the diagram:

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

Key: 4 | 8 = 48

For larger sets of data, a **frequency distribution table** can be drawn. Information is grouped and the *frequency*, the number of times that a particular value or group of values occurs, is stated for each group. For example, the table to the right lists the scores of 250 students on a test administered to all tenth graders in a school district.

The table tells us that 24 students scored between 91 and 100, that 82 scored between 81 and 90, that 77 scored between 71 and 80. The table also tells us that the largest number of students scored in the 80's and that ten students scored 50 or below. Note that unlike the stem-and-leaf diagram, the table does not give us the individual scores in each interval.

Score	Frequency
91–100	24
81–90	82
71–80	77
61–70	36
51–60	21
41–50	8
31–40	2

EXAMPLE 2

The prices of a gallon of milk in 15 stores are listed below.

\$3.15 \$3.39 \$3.28 \$2.98 \$3.25 \$3.45 \$3.58 \$3.24
 \$3.35 \$3.29 \$3.29 \$3.30 \$3.25 \$3.40 \$3.29

- Organize the data in a stem-and-leaf diagram.
- Display the data in a frequency distribution table.
- If the 15 stores were chosen at random from the more than 100 stores that sell milk in Monroe County, does the data set represent a population or a sample?

Solution a. Use the first two digits of the price as the stem.

Use the last digit as the leaf.

Write the leaves in numerical order.

Stem	
3.5	
3.4	
3.3	
3.2	
3.1	
3.0	
2.9	

Stem	Leaf
3.5	8
3.4	5 0
3.3	9 5 0
3.2	8 5 4 9 9 5 9
3.1	5
3.0	
2.9	8

Stem	Leaf
3.5	8
3.4	0 5
3.3	0 5 9
3.2	4 5 5 8 9 9 9
3.1	5
3.0	
2.9	8

Key: 2.9 | 8 = 2.98

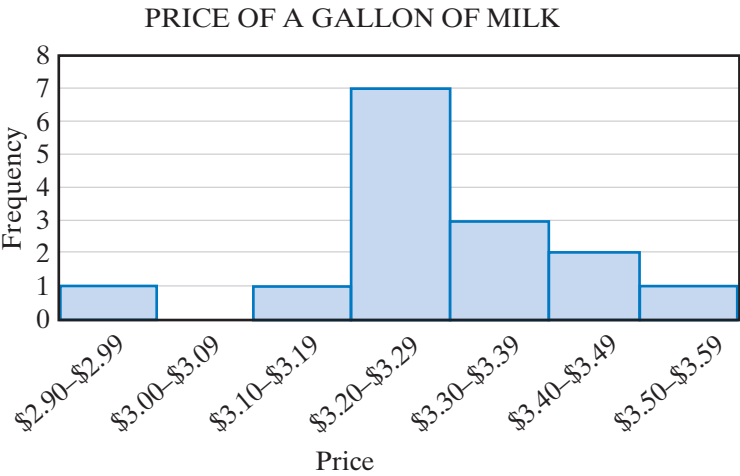
- b. Divide the data into groups of length \$0.10 starting with \$2.90. These groups correspond with the stems of the stem-and-leaf diagram. The frequencies can be determined by the use of a tally to represent each price.

Stem	Leaf	Price	Tally	Frequency
3.5	8	\$3.50–\$3.59		1
3.4	0 5	\$3.40–\$3.49		2
3.3	0 5 9	\$3.30–\$3.39		3
3.2	4 5 5 8 9 9 9	\$3.20–\$3.29		7
3.1	5	\$3.10–\$3.19		1
3.0		\$3.00–\$3.09		0
2.9	8	\$2.90–\$2.99		1

Key: 2.9 | 8 = 2.98

- c. The data set is obtained from a random selection of stores from all of the stores in the study and is therefore a sample. [Answer](#)

Once the data have been organized, a graph can be used to visualize the intervals and their frequencies. A **histogram** is a vertical bar graph where each interval is represented by the width of the bar and the frequency of the interval is represented by the height of the bar. The bars are placed next to each other to show that, as one interval ends, the next interval begins. The histogram below shows the data of Example 2:





A graphing calculator can be used to display a histogram from the data on a frequency distribution table.

- (1) Clear L_1 and L_2 , the lists to be used, of existing data.

ENTER: **STAT** **4** **2nd** **L1** **,** **2nd** **L2** **ENTER**

- (2) Press **STAT** **1** to edit the lists. Enter the *minimum* value of each interval in L_1 and the frequencies into L_2 .

L1	L2	L3
3.5	1	
3.4	2	
3.3	3	
3.2	7	
3.1	1	
3	0	
2.9	1	

L3(1)=

- (3) Clear any functions in the **Y=** menu.

- (4) Turn on Plot1 from the STAT PLOT menu and select for histogram. Make sure to also set *Xlist* to L_1 and *Freq* to L_2 .

ENTER: **2nd** **STAT PLOT** **1** **ENTER**

▼ **▶** **▶** **ENTER** **▼** **2nd**

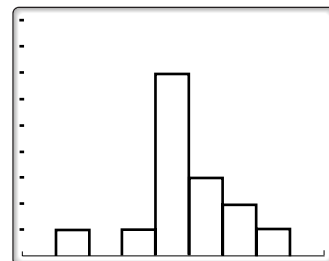
L1 **▼** **2nd** **L2**

Plot1	Plot2	Plot3
On	Off	
Type:		
Xlist:	L1	
Freq:	L2	

- (5) In the **WINDOW** menu, enter X_{min} as 2.8, the length of one interval less than the smallest interval value, and X_{max} as 3.7, the length of one interval more than the largest interval value. Enter X_{scl} as 0.10, the length of the interval. The Y_{min} is 0 and Y_{max} is 9 to be greater than the largest frequency.

WINDOW
$X_{min}=2.8$
$X_{max}=3.7$
$X_{scl}=1$
$Y_{min}=0$
$Y_{max}=9$
$Y_{scl}=1$
$X_{res}=1$

- (6) Press **GRAPH** to graph the histogram. We can view the frequency associated with each interval by pressing **TRACE**. Use the left and right arrows to move from one interval to the next.



Exercises

Writing About Mathematics

1. In a controlled experiment, two groups are formed to determine the effectiveness of a new cold remedy. One group takes the medicine and one does not. Explain why the two groups are necessary.
2. In the experiment described in Exercise 1, explain why it is necessary that a participant does not know to which group he or she belongs.

Developing Skills

In 3–5, organize the data in a stem-and-leaf diagram.

3. The grades on a chemistry test:

95 90 84 85 74 67 78 86 54 82
75 67 92 66 90 68 88 85 76 87

4. The weights of people starting a weight-loss program:

173 210 182 190 175 169 236 192 203 196 201
187 205 195 224 177 195 207 188 184 196 155

5. The heights, in centimeters, of 25 ten-year-old children:

137 134 130 144 131 141 136 140 137 129 139 137 144
127 147 143 132 132 142 142 131 129 138 151 137

In 6–8, organize the data in a frequency distribution table.

6. The numbers of books read during the summer months by each of 25 students:

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

7. The sizes of 26 pairs of jeans sold during a recent sale:

8 12 14 10 12 16 14 6 10 9 8 13 12
8 12 10 12 14 10 12 16 10 11 15 8 14

8. The number of siblings of each of 30 students in a class:

2 1 1 5 1 0 2 2 1 3 4 0 6 2 0
3 1 2 2 1 1 1 0 2 1 0 1 1 2 3

In 9–11, graph the histogram of each set of data.

9.

x_i	f_i
35–39	13
30–34	19
25–29	10
20–24	13
15–19	8
10–14	19
5–9	15

10.

x_i	f_i
101–110	3
91–100	6
81–90	10
71–80	13
61–70	14
51–60	2
41–50	2

11.

x_i	f_i
\$55–\$59	20
\$50–\$54	15
\$45–\$49	12
\$40–\$44	5
\$35–\$39	10
\$30–\$34	12
\$25–\$29	16

Applying Skills

In 12–18, suggest a method that might be used to collect data for each study. Tell whether your method uses a population or a sample.

12. Average temperature for each month for a given city
13. Customer satisfaction at a restaurant
14. Temperature of a patient in a hospital over a period of time
15. Grades for students on a test
16. Population of each of the states of the United States
17. Heights of children entering kindergarten
18. Popularity of a new movie
19. The grades on a math test of 25 students are listed below.

86 92 77 84 75 95 66 88 84 53 98 87 83
74 61 82 93 98 87 77 86 58 72 76 89

- a. Organize the data in a stem-and-leaf diagram.
 - b. Organize the data in a frequency distribution table.
 - c. How many students scored 70 or above on the test?
 - d. How many students scored 60 or below on the test?
20. The stem-and-leaf diagram at the right shows the ages of 30 people in an exercise class. Use the diagram to answer the following questions.
- a. How many people are 45 years old?
 - b. How many people are older than 60?
 - c. How many people are younger than 30?
 - d. What is the age of the oldest person in the class?
 - e. What is the age of the youngest person in the class?

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

Key: 1 | 9 = 19

Hands-On Activity

In this activity, you will take a survey of 25 people. You will need a stopwatch or a clock with a second hand. Perform the following experiment with each participant to determine how each perceives the length of a minute:

1. Indicate the starting point of the minute.
2. Have the person tell you when he or she believes that a minute has passed.
3. Record the actual number of seconds that have passed.

After surveying all 25 participants, use a stem-and-leaf diagram to record your data. Keep this data. You will use it throughout your study of this chapter.

15-2 MEASURES OF CENTRAL TENDENCY

After data have been collected, it is often useful to represent the data by a single value that in some way seems to represent all of the data. This number is called the **measure of central tendency**. The most frequently used measures of central tendency are the *mean*, the *median*, and the *mode*.

The Mean

The **mean** or **arithmetic mean** is the most common measure of central tendency. The mean is the sum of all of the data values divided by the number of data values.

For example, nine members of the basketball team played during all or part of the last game. The number of points scored by each of the players was:

21, 15, 12, 9, 8, 7, 5, 2, 2

$$\text{Mean} = \frac{21 + 15 + 12 + 9 + 8 + 7 + 5 + 2 + 2}{9} = \frac{81}{9} = 9$$

Note that if each of the 9 players had scored 9 points, the total number of points scored in the game would have been the same.

The summation symbol, Σ , is often used to designate the sum of the data values. We designate a data value as x_i and the sum of n data values as

$$\sum_{i=1}^n x_i = x_1 + x_2 + x_3 + \cdots + x_n$$

For the set of data given above:

$$n = 9, x_1 = 21, x_2 = 15, x_3 = 12, x_4 = 9, x_5 = 8, x_6 = 7, x_7 = 5, x_8 = 2, x_9 = 2$$

$$\sum_{i=1}^9 x_i = 81$$

The subscript for each data value indicates its position in a list of data values, not its value, although the value of i and the data value may be the same.

Procedure

To find the mean of a set of data:

1. Add the data values.
2. Divide the sum by n , the total number of data values.

EXAMPLE 1

An English teacher recorded the number of spelling errors in the 40 essays written by students. The table below shows the number of spelling errors and the frequency of that number of errors, that is, the number of essays that contained that number of misspellings. Find the mean number of spelling errors for these essays.

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

Solution To find the total number of spelling errors, first multiply each number of errors by the frequency with which that number of errors occurred. For example, since 2 essays each contained 10 errors, there were 20 errors in these essays. Add the products in the $f_i x_i$ row to find the total number of errors in the essays. Divide this total by the total frequency.

												Total
Errors (x_i)	0	1	2	3	4	5	6	7	8	9	10	
Frequency (f_i)	1	3	2	2	6	9	7	5	2	1	2	40
Errors · Frequency ($f_i x_i$)	0	3	4	6	24	45	42	35	16	9	20	204

$$\text{Mean} = \frac{\sum_{i=0}^{10} f_i x_i}{40} = \frac{204}{40} = 5.1$$

Note that for this set of data, the data value is equal to i for each x_i .

Answer 5.1 errors

The Median

The **median** is the middle number of a data set arranged in numerical order. When the data are arranged in numerical order, the number of values less than or equal to the median is equal to the number of values greater than or equal to

the median. Consider again the nine members of the basketball team who played during all or part of the last game and scored the following number of points:

$$2, 2, 5, 7, 8, 9, 12, 15, 21$$

We can write $9 = 4 + 1 + 4$. Therefore, we think of four scores below the median, the median, and four scores above the median.

$$\underbrace{2, 2, 5, 7}_{\text{scores less than the median}}, \quad \overset{\text{median}}{\underset{\uparrow}{8}}, \quad \underbrace{9, 12, 15, 21}_{\text{scores greater than the median}}$$

Note that the number of data values can be written as $2(4) + 1$. The median is the $(4 + 1)$ or 5th value from either end of the distribution. The median number of points scored is 8.

When the number of values in a set of data is even, then the median is the mean of the two middle values. For example, eight members of a basketball team played in a game and scored the following numbers of points.

$$4, 6, 6, 7, 11, 12, 18, 20$$

We can separate the eight data values into two groups of 4 values. Therefore, we average the largest score of the four lowest scores and the smallest score of the four highest scores.

$$\underbrace{4, 6, 6, 7}_{\text{four lowest scores}}, \quad \overset{\text{median}}{\underset{\uparrow}{\frac{7 + 11}{2}}}, \quad \underbrace{11, 12, 18, 20}_{\text{four highest scores}}$$

$$\text{median} = \frac{7 + 11}{2} = 9$$

Note that the number of data values can be written as $2(4)$. The median is the mean of the 4th and 5th value from either end of the distribution. The median is 9. There are four scores greater than the median and four scores lower than the median. The median is a middle mark.

Procedure

To find the median of a set of data:

1. Arrange the data in order from largest to smallest or from smallest to largest.
2. **a.** If the number of data values is odd, write that number as $2n + 1$. The median is the score that is the $(n + 1)$ th score from either end of the distribution.
- b.** If the number of data values is even, write that number as $2n$. The median is the score that is the mean of the n th score and the $(n + 1)$ th score from either end of the distribution.

The Mode

The **mode** is the value or values that occur most frequently in a set of data. For example, in the set of numbers

3, 4, 5, 5, 6, 6, 6, 6, 7, 7, 8, 10,

the number that occurs most frequently is 6. Therefore, 6 is the mode for this set of data.

When a set of numbers is arranged in a frequency distribution table, the mode is the entry with the highest frequency. The table to the right shows, for a given month, the number of books read by each student in a class. The largest number of students, 12, each read four books. The mode for this distribution is 4.

Number of Books Read	Frequency
8	1
7	1
6	3
5	6
4	12
3	4
2	2
1	0
0	1

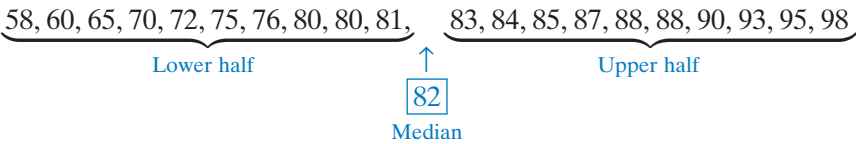
A data set may have more than one mode. For example, in the set of numbers 3, 4, 5, 5, 5, 5, 6, 6, 6, 6, 7, 7, 8, 10, the numbers 5 and 6 each occur four times, more frequently than any other number. Therefore, 5 and 6 are modes for this set of data. The set of data is said to be **bimodal**.

Quartiles

When a set of data is listed in numerical order, the median separates the data into two equal parts. The **quartiles** separate the data into four equal parts. To find the quartiles, we first separate the data into two equal parts and then separate each of these parts into two equal parts. For example, the grades of 20 students on a math test are listed below.

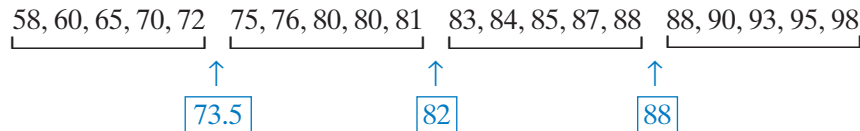
58, 60, 65, 70, 72, 75, 76, 80, 80, 81, 83, 84, 85, 87, 88, 88, 90, 93, 95, 98

Since there are 20 or 2(10) grades, the median grade that separates the data into two equal parts is the average of the 10th and 11th grade.



The 10th grade is 81 and the 11th grade is 83. Therefore, the mean of these two grades, $\frac{81 + 83}{2}$ or 82, separates the data into two equal parts. This number is the median grade.

Now separate each half into two equal parts. Find the median of the two lower quarters and the median of the two upper quarters.



The numbers 73.5, 82, and 88 are the quartiles for this data.

- One quarter of the grades are less than or equal to 73.5. Therefore, 73.5 is the **first quartile** or the **lower quartile**.
- Two quarters of the grades are less than or equal to 82. Therefore, 82 is the **second quartile**. The second quartile is always the median.
- Three quarters of the grades are less than or equal to 88. Therefore, 88 is the **third quartile** or the **upper quartile**.

Note: The minimum, first quartile, median, third quartile, and maximum make up the **five statistical summary** of a data set.

When the data set has an odd number of values, the median or second quartile will be one of the values. This number is not included in either half of the data when finding the first and third quartiles.

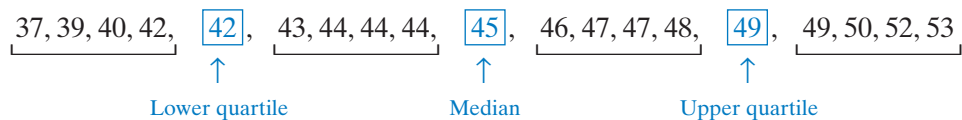
For example, the heights, in inches, of 19 children are given below:

37, 39, 40, 42, 42, 43, 44, 44, 44, 45, 46, 47, 47, 48, 49, 49, 50, 52, 53

There are $19 = 2(9) + 1$ data values. Therefore, the second quartile is the 10th height or 45.



There are $9 = 2(4) + 1$ heights in the lower half of the data and also in the upper half of the data. The middle height of each half is the 5th height.



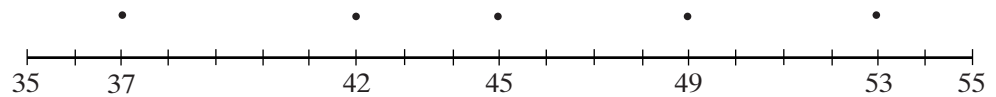
In this example, the first or lower quartile is 42, the median or second quartile is 45, and the third or upper quartile is 49. Each of these values is one of the data values, and the remaining values are separated into four groups with the same number of heights in each group.

Box-and-Whisker Plot

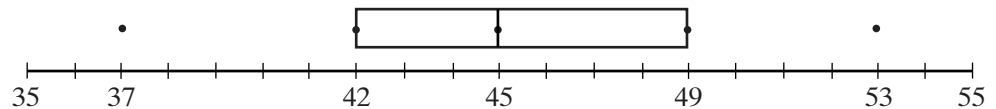
A box-and-whisker plot is a diagram that is used to display the quartile values and the maximum and minimum values of a distribution. We will use the data from the set of heights given above.

37, 39, 40, 42, 42, 43, 44, 44, 44, 45, 46, 47, 47, 48, 49, 49, 50, 52, 53

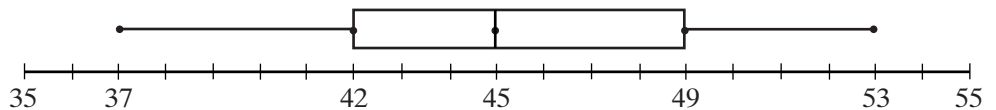
1. Choose a scale that includes the maximum and minimum values of the data. We will use a scale from 35 to 55.
2. Above the scale, place dots to represent the minimum value, the lower quartile, the mean, the upper quartile, and the maximum value.



3. Draw a box with opposite sides through the lower and upper quartiles and a vertical line through the median.



4. Draw whiskers by drawing a line to join the dot that represents the minimum value to the dot that represents the lower quartile and a line to join the dot that represents the upper quartile to the dot that represents the maximum value.



A graphing calculator can be used to find the quartiles and to display the box-and-whisker plot.

- (1) Enter the data for this set of heights into L_1 .
- (2) Turn off all plots and enter the required choices in Plot 1.

ENTER: 2nd STAT PLOT 1 ENTER

▼ ▶ ▶ ▶ ▶ ENTER ▼

2nd L1 ▼ ALPHA 1

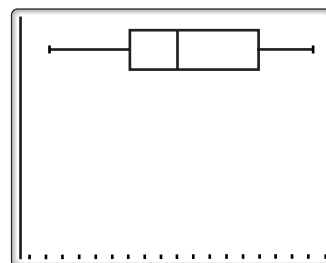


(3) Now display the plot by entering

ZOOM **9**. You can press **TRACE** to display the five statistical summary.

The five statistical summary of a set of data can also be displayed on the calculator by using 1-Var Stats.

ENTER: **STAT** **▶** **ENTER** **ENTER**



The first entry under 1-Var Stats is \bar{x} , the value of the mean. The next is $\sum x$, the sum of the data values. The next three entries are values that will be used in the sections that follow. The last entry is the number of data values. The arrow tells us that there is more information. Scroll down to display the minimum value, $\min X = 37$, the lower quartile, $Q_1 = 42$, the median or second quartile, $Med = 45$, the upper quartile, $Q_3 = 49$, and the maximum value, $\max X = 53$.

```
1-VAR STATS
 $\bar{x}$ =45.31578947
 $\sum x$ =861
 $\sum x^2$ =39353
 $s_x$ =4.321170515
 $\sigma x$ =4.205918531
↓n=19
```

```
1-VAR STATS
↑n=19
minX=37
Q1=42
Med=45
Q3=49
maxX=53
```

EXAMPLE 1

Find the mean, the median, and the mode of the following set of grades:

92, 90, 90, 90, 88, 87, 85, 70

Solution Mean = $\frac{1}{n} \sum_{i=1}^n x_i = \frac{1}{8}(92 + 90 + 90 + 90 + 88 + 87 + 85 + 70) = \frac{692}{8} = 86.5$

Median = the average of the 4th and 5th grades = $\frac{90 + 88}{2} = 89$

Mode = the grade that appears most frequently = 90

EXAMPLE 2

The following list shows the length of time, in minutes, for each of 35 employees to commute to work.

25 12 20 18 35 25 40 35 27 30 60 22 36 20 18
27 35 42 35 55 27 30 15 22 10 35 27 15 57 18
25 45 24 27 25

- Organize the data in a stem-and-leaf diagram.
- Find the median, lower quartile, and upper quartile.
- Draw a box-and-whisker plot.

Solution a. (1) Choose the tens digit as the stem and enter the units digit as the leaf for each value.

Stem	Leaf
6	0
5	5 7
4	0 2 5
3	5 5 0 6 5 5 0 5
2	5 0 5 7 2 0 7 7 2 7 5 4 7 5
1	2 8 8 5 0 5 8

- (2) Write the leaves in numerical order from smallest to largest.

Stem	Leaf
6	0
5	5 7
4	0 2 5
3	0 0 5 5 5 5 5 6
2	0 0 2 2 4 5 5 5 5 7 7 7 7 7
1	0 2 5 5 8 8 8

Key: 1 | 0 = 10

- b. (1) The median is the middle value of the 35 data values when the values are arranged in order.

$$35 = 2(17) + 1$$

The median is the 18th value. Separate the data into groups of 17 from each end of the distribution. The median is 27.

Stem	Leaf
6	0
5	5 7
4	0 2 5
3	0 0 5 5 5 5 5 6
2	0 0 2 2 4 5 5 5 5 7 7 7 7 7
1	0 2 5 5 8 8 8

- (2) There are $17 = 2(8) + 1$ data values below the median. The lower quartile is the 9th data value from the lower end. The upper quartile is the 9th data value from the upper end.

Stem	Leaf
6	0
5	5 7
4	0 2 5
3	0 0 5 5 5 5 5 6
2	0 0 2 2 4 5 5 5 5 7 7 7 7 7
1	0 2 5 5 8 8 8

The lower quartile is 20, the median is 27, and the upper quartile is 35.



Note: In the stem-and-leaf diagram of Example 2, the list is read from left to right to find the lower quartile. The list is read from *right* to *left* (starting from the top) to find the upper quartile.

Exercises

Writing About Mathematics

1. Cameron said that the number of data values of any set of data that are less than the lower quartile or greater than the upper quartile is exactly 50% of the number of data values. Do you agree with Cameron? Explain why or why not.
2. Carlos said that for a set of $2n$ data values or of $2n + 1$ data values, the lower quartile is the median of the smallest n values and the upper quartile is the median of the largest n values. Do you agree with Carlos? Explain why or why not.

Developing Skills

In 3–8, find the mean, the median, and the mode of each set of data.

3. Grades: 74, 78, 78, 80, 80, 80, 82, 88, 90
4. Heights: 60, 62, 63, 63, 64, 65, 66, 68, 68, 68, 70, 75
5. Weights: 110, 112, 113, 115, 115, 116, 118, 118, 125, 134, 145, 148
6. Number of student absences: 0, 0, 0, 1, 1, 2, 2, 2, 3, 4, 5, 9
7. Hourly wages: \$6.90, \$7.10, \$7.50, \$7.50, \$8.25, \$9.30, \$9.50, \$10.00
8. Tips: \$1.00, \$1.50, \$2.25, \$3.00, \$3.30, \$3.50, \$4.00, \$4.75, \$5.00, \$5.00, \$5.00

In 9–14, find the median and the first and third quartiles for each set of data values.

9. 2, 3, 5, 8, 9, 11, 15, 16, 17, 20, 22, 23, 25
10. 34, 35, 35, 36, 38, 40, 42, 43, 43, 43, 44, 46, 48, 50
11. 23, 27, 15, 38, 12, 17, 22, 39, 28, 20, 27, 18, 25, 28, 30, 29
12. 92, 86, 77, 85, 88, 90, 81, 83, 95, 76, 65, 88, 91, 81, 88, 87, 95
13. 75, 72, 69, 68, 66, 65, 64, 63, 63, 61, 60, 59, 59, 58, 56, 54, 52, 50
14. 32, 32, 30, 30, 29, 27, 26, 22, 20, 20, 19, 18, 17
15. A student received the following grades on six tests: 90, 92, 92, 95, 95, x .
 - a. For what value(s) of x will the set of grades have no mode?
 - b. For what value(s) of x will the set of grades have only one mode?
 - c. For what value(s) of x will the set of grades be bimodal?
16. What are the first, second, and third quartiles for the set of integers from 1 to 100?
17. What are the first, second, and third quartiles for the set of integers from 0 to 100?

Applying Skills

18. The grades on an English test are shown in the stem-and-leaf diagram to the right.
- Find the mean grade.
 - Find the median grade.
 - Find the first and third quartiles.
 - Draw a box-and-whisker plot for this data.

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

Key: 4 | 7 = 47

19. The weights in pounds of the members of the football team are shown below:

181 199 178 203 211 208 209 202 212 194
 185 208 223 206 202 213 202 186 189 203

- Find the mean.
 - Find the median.
 - Find the mode or modes.
 - Find the first and third quartiles.
 - Draw a box-and-whisker plot.
20. Mrs. Gillis gave a test to her two classes of algebra. The mean grade for her class of 20 students was 86 and the mean grade of her class of 15 students was 79. What is the mean grade when she combines the grades of both classes?

Hands-On Activity

Use the estimates of a minute collected in the Hands-On Activity of the previous section to determine the five statistical summary for your data. Draw a box-and-whisker plot to display the data.

I 5-3 MEASURES OF CENTRAL TENDENCY FOR GROUPED DATA

Most statistical studies involve much larger numbers of data values than can be conveniently displayed in a list showing each data value. Large sets of data are usually organized into a frequency distribution table.

Frequency Distribution Tables for Individual Data Values

A frequency distribution table records the individual data values and the frequency or number of times that the data value occurs in the data set. The example on page 606 illustrates this method of recording data.

Each of an English teacher's 100 students recently completed a book report. The teacher recorded the number of misspelled words in each report. The table

records the number of reports for each number of misspelled words. Let x_i represent the number of misspelled words in a report and f_i represent the number of reports that contain x_i misspelled words.

	Total										
x_i	0	1	2	3	4	5	6	7	8	9	10
f_i	5	7	6	8	19	26	16	7	5	0	1
$x_i f_i$	0	7	12	24	76	130	96	49	40	0	10

The mean of this set of data is the total number of misspelled words divided by the number of reports. To find the total number of misspelled words, we must first multiply each number of misspelled words, x_i , by the number of reports that contain that number of misspelled words, f_i . That is, we must find $x_i f_i$ for each number of misspelled words.

For this set of data, when we sum the f_i row, $\sum_{i=0}^{10} f_i = 100$, and when we sum the $x_i f_i$ row $\sum_{i=0}^{10} x_i f_i = 444$.

$$\text{Mean} = \frac{\sum_{i=0}^{10} x_i f_i}{\sum_{i=0}^{10} f_i} = \frac{444}{100} = 4.44$$

To find the median and the quartiles for this set of data, we will find the *cumulative frequency* for each number of misspelled words. The **cumulative frequency** is the accumulation or the sum of all frequencies less than or equal to a given frequency. For example, the cumulative frequency for 3 misspelled words on a report is the sum of the frequencies for 3 or fewer misspelled words, and the cumulative frequency for 6 misspelled words on a report is the sum of the frequencies for 6 or fewer misspelled words. The third row of the table shows that 0 misspelled words occur 5 times, 1 or fewer occur $7 + 5$ or 12 times, 2 or fewer occur $6 + 12$ or 18 times. In each case, the cumulative frequency for x_i is the frequency for x_i plus the cumulative frequency for x_{i-1} . The cumulative frequency for the largest data value is always equal to the total number of data values.

x_i	0	1	2	3	4	5	6	7	8	9	10
f_i	5	7	6	8	19	26	16	7	5	0	1
Cumulative Frequency	5	12	18	26	45	71	87	94	99	99	100

↑
3
Lower quartile

↑
5
Median

↑
6
Upper quartile

Since there are 100 data values, the median is the average of the 50th and 51st values. To find these values, look at the cumulative frequency column. There are 45 values less than or equal to 4 and 71 less than or equal to 5. Therefore, the 50th and 51st values are both 5 and the median is 5 misspelled words.

Similarly, the upper quartile is the average of the 75th and 76th values. Since there are 71 values less than or equal to 5, the 75th and 76th values are both 6 misspelled words. The lower quartile is the average of the 25th and 26th values. Since there are 18 values less than or equal to 2, the 25th and 26th values are both 3 misspelled words.

Percentiles

A **percentile** is a number that tells us what percent of the total number of data values lie at or below a given measure.

For example, let us use the data from the previous section. The table records the number of reports, f_i , that contain each number of misspelled words, x_i , on 100 essays.

x_i	0	1	2	3	4	5	6	7	8	9	10
f_i	5	7	6	8	19	26	16	7	5	0	1
Cumulative Frequency	5	12	18	26	45	71	87	94	99	99	100

To find the percentile rank of 7 misspelled words, first find the number of essays with fewer than 7 misspelled words, 87. Add to this *half* of the essays with 7 misspelled words, $\frac{7}{2}$ or 3.5. Add these two numbers and divide the sum by the number of essays, 100.

$$\frac{87 + 3.5}{100} = \frac{90.5}{100} = 90.5\%$$

Percentiles are usually not written with fractions. We say that 7 misspelled words is at the 90.5th or 91st percentile. That is, 91% of the essays had 7 or fewer misspelled words.

Frequency Distribution Tables for Grouped Data

Often the number of different data values in a set of data is too large to list each data value separately in a frequency distribution table. In this case, it is useful to list the data in terms of groups of data values rather than in terms of individual data values. The following example illustrates this method of recording data.

There are 50 members of a weight-loss program. The weights range from 181 to 285 pounds. It is convenient to arrange these weights in groups of 10 pounds starting with 180–189 and ending with 280–289. The frequency distribution table shows the frequencies of the weights for each interval.

Weights	Midpoint x_i	Frequency f_i	$x_i f_i$	Cumulative Frequency
280–289	284.5	1	284.5	50
270–279	274.5	3	823.5	49
260–269	264.5	4	1,058.0	46
250–259	254.5	8	2,036.0	42
240–249	244.5	12	2,934.0	34
230–239	234.5	10	2,345.0	22
220–229	224.5	5	1,122.5	12
210–219	214.5	3	643.5	7
200–209	204.5	2	409.0	4
190–199	194.5	1	194.5	2
180–189	184.5	1	184.5	1
		50	12,035	

In order to find the mean, assume that the weights are evenly distributed throughout the interval. The mean is found by using the midpoint of the weight intervals as representative of each value in the interval groupings.

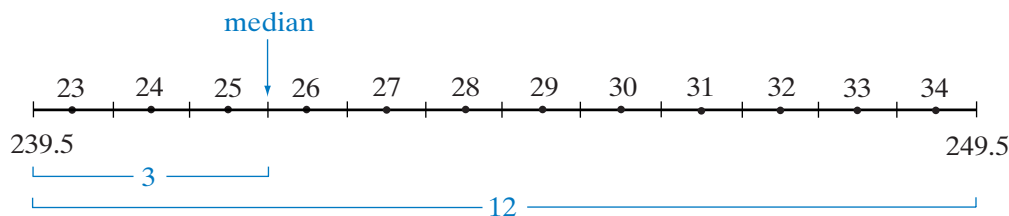
$$\text{Mean} = \frac{\sum_{i=1}^{11} x_i f_i}{\sum_{i=1}^{11} f_i} = \frac{12,035}{50} = 240.7$$

To find the median for a set of data that is organized in intervals greater than 1, first find the interval in which the median lies by using the cumulative frequency.

There are 50 data values. Therefore, the median is the value between the 25th and the 26th values. The cumulative frequency tells us that there are 22 values less than or equal to 239 and 34 values less than or equal to 249. Therefore, the 25th and 26th values are in the interval 240–249. Can we give a better approximation for the median?

The endpoints of an interval are the lowest and highest data values to be entered in that interval. The boundary values are the values that separate intervals. The lower boundary of an interval is midway between the lower endpoint of the interval and the upper endpoint of the interval that precedes it. The lower boundary of the 240–249 is midway between 240 and 239, that is, 239.5. The upper endpoint of this interval is between 249 and 250, that is 249.5.

Since there are 34 weights less than or equal to 249 and 22 weights less than 240, the weights in the 240–249 interval are the 23rd through the 34th weights. Think of these 12 weights as being evenly spaced throughout the interval.



The midpoint between the 25th and 26th weights is $\frac{3}{12}$ of the distance between the boundaries of the interval, a difference of 10.

$$\text{Median} = 239.5 + \frac{3}{12}(10) = 239.5 + 2.5 = 242$$

The estimated median for the weights is 242 pounds. Thus, if we assume that the data values are evenly distributed within each interval, we can obtain a better approximation for the median.

EXAMPLE I

The numbers of pets owned by the children in a sixth-grade class are given in the table.

- Find the mean.
- Find the median.
- Find the mode for this set of data.
- Find the percentile rank of 4 pets.

No. of Pets	Frequency
6	2
5	1
4	3
3	5
2	8
1	14
0	7

Solution a. Add to the table the number of pets times the frequency, and the cumulative frequency for each row of the table.

No. of Pets x_i	Frequency f_i	$f_i x_i$	Cumulative Frequency
6	2	12	40
5	1	5	38
4	3	12	37
3	5	15	34
2	8	16	29
1	14	14	21
0	7	0	7
	40	74	

Add the numbers in the $f_i x_i$ column and the numbers in the f_i column.

$$\sum_{i=0}^6 f_i = 40 \qquad \sum_{i=0}^6 f_i x_i = 74$$

$$\text{Mean} = \frac{\sum_{i=0}^6 x_i f_i}{\sum_{i=0}^6 f_i} = \frac{74}{40} = 1.85$$

- There are 40 data values in this set of data. The median is the average of the 20th and the 21st data values. Since there are 21 students who own 1 or fewer pets, both the 20th and the 21st data value is 1. Therefore, the median number of pets is 1.
- The largest number of students, 14, have 1 pet. The mode is 1.
- There are 34 students with fewer than 4 pets and 3 students with 4 pets. Add 34 and half of 3 and divide the sum by the total number of students, 40.

$$\frac{34 + \frac{3}{2}}{40} = \frac{35.5}{40} = 0.8875 = 88.75\%$$

Four pets represents the 89th percentile.

**Calculator
Solution for
a, b, and c**

Clear the lists first if necessary by keying in **STAT** **4** **2nd** **L1** **,** **2nd** **L2** **ENTER**.

Enter the number of pets in L_1 .

Enter the frequency for each number of pets in L_2 .

Display 1-Var Stats.

ENTER: **STAT** **►** **1** **2nd** **L1** **,** **2nd** **L2** **ENTER**

DISPLAY:

```
1-VAR STATS
x̄=1.85
Σx=74
Σx²=236
sx=1.594059485
σx=1.574007624
↓N=40
```

```
1-VAR STATS
↑N=40
MINX=0
Q1=1
MED=1
Q3=3
MAXX=6
```

The first entry, \bar{x} , is the mean, 1.85. Use the down-arrow key, **▼**, to display the median, *Med*. The median is 1.

Note that $\sum x = 74$ is $\sum_{i=0}^6 f_i x_i$ and $n = 40$ is $\sum_{i=0}^6 f_i$.

Answers **a.** mean = 1.85 **b.** median = 1 **c.** mode = 1 **d.** 4 is the 89th percentile

Note: The other information displayed under 1-Var Stats will be used in the sections that follow.

EXAMPLE 2

A local business made the summary of the ages of 45 employees shown below. Find the mean and the median age of the employees to the nearest integer.

Age	20–24	25–29	30–34	35–39	40–44	45–49	50–54	55–59	60–64
Frequency	1	2	5	6	7	10	7	5	2

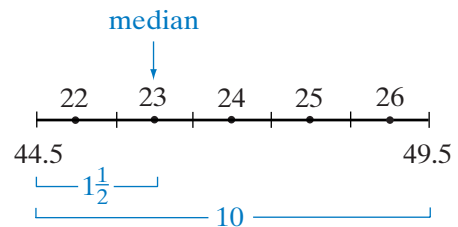
Solution Add to the table the midpoint, the midpoint times the frequency, and the cumulative frequency for each interval.

Age	Midpoint (x_i)	Frequency (f_i)	$x_i f_i$	Cumulative Frequency
60–64	62	2	124	45
55–59	57	5	285	43
50–54	52	7	364	38
45–49	47	10	470	31
40–44	42	7	294	21
35–39	37	6	222	14
30–34	32	5	160	8
25–29	27	2	54	3
20–24	22	1	22	1
		45	1,995	

$$\text{Mean} = \frac{\sum_{i=1}^{11} x_i f_i}{\sum_{i=1}^{11} f_i} = \frac{1,995}{45} = 44.333 \dots \approx 44$$

The mean is the middle age of 45 ages or the 23rd age. Since there are 21 ages less than 45, the 23rd age is the second of 10 ages in the 45–49 interval. The boundaries of the 45–49 interval are 44.5 to 49.5

$$\begin{aligned} \text{Median} &= 44.5 + \frac{1}{2} \times 5 \\ &= 44.5 + 0.75 = 45.25 \approx 45 \end{aligned}$$



Answer mean = 44, median = 45

Exercises

Writing About Mathematics

- Adelaide said that since, in Example 2, there are 10 employees whose ages are in the 45–49 interval, there must be two employees of age 45. Do you agree with Adelaide? Explain why or why not.
- Gail said that since, in Example 2, there are 10 employees whose ages are in the 45–49 interval, there must at least two employees who are the same age. Do you agree with Gail? Explain why or why not.

Developing Skills

In 3–8, find the mean, the median, and the mode for each set of data.

3.

x_i	f_i
5	6
4	10
3	15
2	11
1	2
0	1

4.

x_i	f_i
50	8
40	12
30	17
20	10
10	3

5.

x_i	f_i
12	7
11	15
10	13
9	16
8	14
7	15
6	9
5	2

6.

x_i	f_i
10	1
9	1
8	3
7	7
6	6
5	2
4	2

7.

x_i	f_i
\$1.10	1
\$1.20	5
\$1.30	8
\$1.40	6
\$1.50	6

8.

x_i	f_i
95	2
90	8
85	12
80	10
75	9
70	3
65	0
60	1

9. Find the percentile rank of 2 for the data in Exercise 3.
10. Find the percentile rank of 20 for the data in Exercise 4.
11. Find the percentile rank of 8 for the data in Exercise 5.
12. Find the percentile rank of 6 for the data in Exercise 6.

In 13–18, find the mean and the median for each set of data to the nearest tenth.

13.

x_i	f_i
21–25	2
16–20	3
11–15	12
6–10	6
1–5	1

14.

x_i	f_i
91–100	5
81–90	8
71–80	10
61–70	6
51–60	0
41–50	1

15.

x_i	f_i
\$1.51–\$1.60	2
\$1.41–\$1.50	5
\$1.31–\$1.40	14
\$1.21–\$1.30	4
\$1.11–\$1.20	2
\$1.01–\$1.10	3

16.

x_i	f_i
17–19	20
14–16	27
11–13	32
8–10	39
5–7	32

17.

x_i	f_i
\$60–\$69	16
\$50–\$59	5
\$40–\$49	16
\$30–\$39	2
\$20–\$29	5
\$10–\$19	37

18.

x_i	f_i
0.151–0.160	16
0.141–0.150	5
0.131–0.140	0
0.121–0.130	6
0.111–0.120	0

Applying Skills

19. The table shows the number of correct answers on a test consisting of 15 questions. Find the mean, the median, and the mode for the number of correct answers.

Correct Answers	6	7	8	9	10	11	12	13	14	15
Frequency	1	0	1	3	5	8	9	6	5	2

20. The ages of students in a calculus class at a high school are shown in the table. Find the mean and median age.

Age	Frequency
19	2
18	8
17	9
16	1
15	1

21. Each time Mrs. Taggart fills the tank of her car, she estimates, from the number of miles driven and the number of gallons of gasoline needed to fill the tank, the fuel efficiency of her car, that is, the number of miles per gallon. The table shows the result of the last 20 times that she filled the car.

- a. Find the mean and the median fuel efficiency (miles per gallon) for her car.
b. Find the percentile rank of 34 miles per gallon.

Miles per Gallon	32	33	34	35	36	37	38	39	40
Frequency	1	3	2	5	3	3	2	0	1

22. The table shows the initial weights of people enrolled in a weight-loss program. Find the mean and median weight.

Weight	191–200	201–210	211–220	221–230	231–240	241–250
Frequency	1	1	5	7	10	12
Weight	251–260	261–270	271–280	281–290	291–300	
Frequency	13	16	8	5	2	

23. In order to improve customer relations, an auto-insurance company surveyed 100 people to determine the length of time needed to complete a report form following an auto accident. The result of the survey is summarized in the following table showing the number of minutes needed to complete the form. Find the mean and median amount of time needed to complete the form.

Minutes	26–30	31–35	36–40	41–45	46–50	51–55	56–60	61–65	66–70
Frequency	2	8	12	15	10	24	26	1	2

Hands-On Activity

Organize the data from the survey in the Hands-On Activity of Section 15-1 using intervals of five seconds. Use the table to find the mean number of seconds. Compare this result with the mean found using the individual data values.

I 5-4 MEASURES OF DISPERSION

The mean and the median of a set of data help us to describe a set of data. However, the mean and the median do not always give us enough information to draw meaningful conclusions about the data. For example, consider the following sets of data.

Ages of students on a middle-school basketball team	Ages of students in a community center tutoring program
11 12 12 13 13 13	6 8 9 9 10 13
13 13 14 14 14 14	13 15 17 18 19 19

The mean of both sets of data is 13 and the median of both sets of data is 13, but the two sets of data are quite different. We need a measure that indicates how the individual data values are scattered or spread on either side of the mean. A number that indicates the variation of the data values about the mean is called a **measure of dispersion**.

Range

The simplest of the measures of dispersion is called the *range*. The **range** is the difference between the highest value and the lowest value of a set of data. In the sets of data given above, the range of ages of students on the middle-school basketball team is $14 - 11$ or 3 and the range of the ages of the students in the community center tutoring program is $19 - 6$ or 13. The difference in the ranges indicates that the ages of the students on the basketball team are more closely grouped about the mean than the ages of the students in the tutoring program.

The range is dependent on only the largest data value and the smallest data value. Therefore, the range can be very misleading. For example consider the following sets of data:

Ages of members of the chess club: 11, 11, 11, 11, 15, 19, 19, 19, 19
 Ages of the members of the math club: 11, 12, 13, 14, 15, 16, 17, 18, 19

For each of these sets of data, the mean is 15, the median is 15, and the range is 8. But the sets of data are very different. The range often does not tell us critical information about a set of data.

Interquartile Range

Another measure of dispersion depends on the first and third quartiles of a distribution. The difference between the first and third quartile values is the **interquartile range**. The interquartile range tells us the range of at least 50% of the data. The largest and smallest values of a set of data are often not representative of the rest of the data. The interquartile range better represents the spread of the data. It also gives us a measure for identifying extreme data values, that is, those that differ significantly from the rest of the data. For example, consider the ages of the members of a book club.

21, 24, 25, 27, 28, 31, 35, 35, 37, 39, 40, 41, 69
 ↑ ↑ ↑
 26 35 39.5
 Q_1 median Q_3

For these 13 data values, the median is the 7th value. For the six values below the median, the first quartile is the average of the 3rd and 4th values from the lower end: $\frac{25 + 27}{2} = 26$. The third quartile is the average of the 3rd and 4th values from the upper end: $\frac{39 + 40}{2} = 39.5$. The interquartile range of the ages of the members of the book club is $39.5 - 26$ or 13.5. The age of the oldest member of the club differs significantly from the ages of the others. It is more than 1.5 times the interquartile range above the upper quartile:

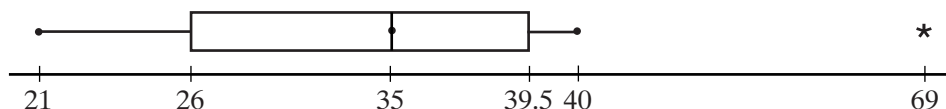
$$69 > 39.5 + 1.5(13.5)$$

We call this data value an *outlier*.

DEFINITION

An **outlier** is a data value that is greater than the upper quartile plus 1.5 times the interquartile range or less than the lower quartile minus 1.5 times the interquartile range.

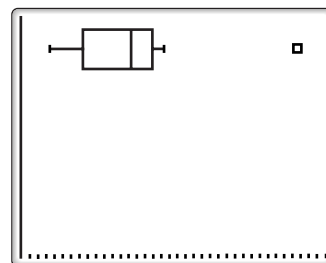
When we draw the box-and-whisker plot for a set of data, the outlier is indicated by a * and the whisker is drawn to the largest or smallest data value that is not an outlier. The box-and-whisker plot for the ages of the members of the book club is shown on page 616.



We can use the graphing calculator to graph a box-and-whisker plot with outliers. From the STAT PLOT menu, choose the option. For example, with the book club data entered into L_1 , the following keystrokes will graph a box-and-whisker plot with outliers.

ENTER: **2nd** **STAT PLOT** **1**
ENTER **▼** **▶** **▶** **▶**
ENTER **▼** **2nd** **L1**
▼ **ALPHA** **1** **ZOOM** **9**

DISPLAY:



EXAMPLE I

The table shows the number of minutes, rounded to the nearest 5 minutes, needed for each of 100 people to complete a survey.

- Find the range and the interquartile range for this set of data.
- Does this data set include outliers?

Minutes	30	35	40	45	50	55	60	65	70	85
Frequency	3	8	12	15	10	24	17	8	2	1
Cumulative Frequency	3	11	23	38	48	72	89	97	99	100

Solution a. The range is the difference between the largest and smallest data value.

$$\text{Range} = 85 - 30 = 55$$

To find the interquartile range, we must first find the median and the lower and upper quartiles. Since there are 100 values, the median is the average of the 50th and the 51st values. Both of these values lie in the interval 55. Therefore, the median is 55. There are 50 values above the mean and 50 values below the mean. Of the lower 50 values, the lower quartile is the average of the 25th and 26th values. Both of these values lie in the interval 45. The lower quartile is 45. The upper quartile is the average of the 25th and the 26th from the upper end of the distribution (or the 75th and 76th from the lower end). These values lie in the interval 60. The upper quartile is 60.

$$\text{Interquartile range} = 60 - 45 = 15$$

- b.** An outlier is a data value that is 1.5 times the interquartile range below the first quartile or above the third quartile.

$$45 - 1.5(15) = 22.5 \qquad 60 + 1.5(15) = 82.5$$

The data value 85 is an outlier.

Answers **a.** Range = 55, interquartile range = 15 **b.** The data value 85 is an outlier.

Exercises

Writing About Mathematics

- In any set of data, is it always true that $x_i = i$? For example, in a set of data with more than three data values, does $x_4 = 4$? Justify your answer.
- In a set of data, $Q_1 = 12$ and $Q_3 = 18$. Is a data value equal to 2 an outlier? Explain why or why not.

Developing Skills

In 3–6, find the range and the interquartile range for each set of data.

3. 3, 5, 7, 9, 11, 13, 15, 17, 19

4. 12, 12, 14, 14, 16, 18, 20, 22, 28, 34

5. 12, 17, 23, 31, 46, 54, 67, 76, 81, 93

6. 2, 14, 33, 34, 34, 34, 35, 36, 37, 37, 38, 40, 42

In 7–9, find the mean, median, range, and interquartile range for each set of data to the nearest tenth.

7.

x_i	f_i
50	3
45	8
40	12
35	15
30	11
25	7
20	4

8.

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

9.

x_i	f_i
11	5
16	8
19	9
31	6
37	5
32	5
35	6

Applying Skills

10. The following data represents the yearly salaries, in thousands of dollars, of 10 basketball players.

533 427 800 687 264 264 125 602 249 19,014

- a. Find the mean and median salaries of the 10 players.
- b. Which measure of central tendency is more representative of the data? Explain.
- c. Find the outlier for the set of data.
- d. Remove the outlier from the set of data and recalculate the mean and median salaries.
- e. After removing the outlier from the set of data, is the mean more or less representative of the data?

11. The grades on a math test are shown in the stem-and-leaf diagram to the right.

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

- a. Find the mean grade.
- b. Find the median grade.
- c. Find the first and third quartiles.
- d. Find the range.
- e. Find the interquartile range.

12. The ages of students in a Spanish class are shown in the table. Find the range and the interquartile range.

Age	Frequency
19	1
18	8
17	8
16	6
15	2

13. The table shows the number of hours that 40 third graders reported studying a week. Find the range and the interquartile range.

Hours	3	4	5	6	7	8	9	10	11	12
Frequency	2	1	3	3	5	8	8	5	4	1

14. The table shows the number of pounds lost during the first month by people enrolled in a weight-loss program.

- a. Find the range.
- b. Find the interquartile range.
- c. Which of the data values is an outlier?

Pounds Lost	1	2	3	4	5	6	7	8	9	11	15
Frequency	1	1	2	2	6	10	7	7	2	1	1

15. The 14 students on the track team recorded the following number of seconds as their best time for the 100-yard dash:

13.5 13.7 13.1 13.0 13.3 13.2 13.0
12.8 13.4 13.3 13.1 12.7 13.2 13.5

Find the range and the interquartile range.

16. The following data represent the waiting times, in minutes, at Post Office A and Post Office B at noon for a period of several days.

A: 1, 2, 2, 2, 3, 3, 3, 3, 3, 3, 9, 10
B: 1, 2, 2, 3, 3, 5, 5, 6, 6, 7, 8, 8, 9, 10

- Find the range of each set of data. Are the ranges the same?
- Graph the box-and-whisker plot of each set of data.
- Find the interquartile range of each set of data.
- If the data values are representative of the waiting times at each post office, which post office should you go to at noon if you are in a hurry? Explain.

Hands-On Activity

Find the range and the interquartile range of the data from the survey in the Hands-On Activity of Section 15-1 estimating the length of a minute. Does your data contain an outlier?

15-5 VARIANCE AND STANDARD DEVIATION

Variance

Let us consider a more significant measure of dispersion than either the range or the interquartile range. Let x_i represent a student's grades on eight tests.

Grade (x_i)	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
95	9	81
92	6	36
88	2	4
87	1	1
86	0	0
82	-4	16
80	-6	36
78	-8	64
$\sum_{i=1}^8 x_i = 688$	$\sum_{i=1}^8 (x_i - \bar{x})^2 = 0$	$\sum_{i=1}^8 (x_i - \bar{x})^2 = 238$

$$\text{Mean} = \frac{\sum_{i=1}^8 x_i}{8} = \frac{688}{8} = 86$$

The table shows the deviation, or difference, of each grade from the mean. Grades above the mean are positive and grades below the mean are negative. For any set of data, the sum of these differences is always 0.

In order to find a meaningful sum, we can use the squares of the differences so that each value will be positive. The square of the deviation of each data value from the mean is used to find another measure of dispersion called the **variance**. To find the variance, find the sum of the squares of the deviations from the mean and divide that sum by the number of data values. In symbols, the variance for a set of data that represents the entire population is given by the formula:

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

For the set of data given above, the variance is $\frac{1}{8}(238)$ or 29.75.

Note that since the *square* of the differences is involved, this method of finding a measure of dispersion gives greater weight to measures that are farther from the mean.

EXAMPLE I

A student received the following grades on five math tests: 84, 97, 92, 88, 79. Find the variance for the set of grades of the five tests.

Solution The mean of this set of grades is:

$$\frac{84 + 97 + 92 + 88 + 79}{5} = \frac{440}{5} = 88$$

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
84	-4	16
97	9	81
92	4	16
88	0	0
79	-9	81
$\sum_{i=1}^5 x_i = 440$		$\sum_{i=1}^5 (x_i - \bar{x})^2 = 194$

$$\text{Variance} = \frac{1}{n} \sum_{i=1}^5 (x_i - \bar{x})^2 = \left(\frac{1}{5}\right) 194 = 38\frac{4}{5} = 38.8 \quad \text{Answer}$$

When the data representing a population is listed in a frequency distribution table, we can use the following formula to find the variance:

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

EXAMPLE 2

In a city, there are 50 math teachers who are under the age of 30. The table below shows the number of years of experience of these teachers. Find the variance of this set of data. Let x_i represent the number of years of experience and f_i represent the frequency for that number of years.

x_i	0	1	2	3	4	5	6	7	8	
f_i	1	2	7	8	6	9	8	4	5	50
$x_i f_i$	0	2	14	24	24	45	48	28	40	225

Solution

$$\text{Mean} = \frac{\sum_{i=0}^8 x_i f_i}{\sum_{i=0}^8 f_i} = \frac{225}{50} = 4.5$$

For this set of data, the data value is equal to i for each x_i .

The table below shows the deviation from the mean, the square of the deviation from the mean, and the square of the deviation from the mean multiplied by the frequency.

x_i	f_i	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	$f_i(x_i - \bar{x})^2$
8	5	3.5	12.25	61.25
7	4	2.5	6.25	25.00
6	8	1.5	2.25	18.00
5	9	0.5	0.25	2.25
4	6	-0.5	0.25	1.50
3	8	-1.5	2.25	18.00
2	7	-2.5	6.25	43.75
1	2	-3.5	12.25	24.50
0	1	-4.5	20.25	20.25
	$\sum_{i=1}^8 f_i = 50$			$\sum_{i=0}^8 f_i(x_i - \bar{x})^2 = 214.50$

$$\text{Variance} = \frac{\sum_{i=0}^8 f_i(x_i - \bar{x})^2}{\sum_{i=0}^8 f_i} = \frac{214.50}{50} = 4.29 \quad \text{Answer}$$

Note that the data given in this example represents a population, that is, data for all of the teachers under consideration.

Standard Deviation Based on the Population

Although the variance is a useful measure of dispersion, it is in square units. For example, if the data were a set of measures in centimeters, the variance would be in square centimeters. In order to have a measure that is in the same unit of measure as the given data, we find the square root of the variance. The square root of the variance is called the **standard deviation**. When the data represents a population, that is, all members of the group being studied:

$$\text{Standard deviation based on a population} = \sqrt{\frac{1}{n} \sum_{i=1}^n f_i (x_i - \bar{x})^2}$$

If the data is grouped in terms of the frequency of a given value:

$$\text{Standard deviation based on a population} = \sqrt{\frac{\sum_{i=1}^n f_i (x_i - \bar{x})^2}{\sum_{i=1}^n f_i}}$$

The symbol for the standard deviation for a set of data that represents a population is σ (lowercase Greek sigma). Many calculators use the symbol σ_x .

EXAMPLE 3

Find the standard deviation for the number of years of experience for the 50 teachers given in Example 2.

Solution The standard deviation is the square root of the variance.

Therefore:

$$\text{Standard deviation} = \sqrt{4.29} = 2.071231518$$

Calculator Solution

- (1) Clear lists 1 and 2.
- (2) Enter the number of years of experience in L_1 and the frequency in L_2 .
- (3) Locate the standard deviation for a population (σ_x) under 1-Var Stats.

ENTER: **STAT** **►** **ENTER**

2nd **L1** **,**

2nd **L2** **ENTER**

DISPLAY:

```
1-VAR STATS
 $\bar{x}$ =4.5
 $\Sigma x$ =225
 $\Sigma x^2$ =1227
 $s_x$ =2.092259788
 $\sigma_x$ =2.071231518
 $\downarrow$ 
 $n$ =50
```

- (4) $\sigma_x = 2.071231518$

Answer The standard deviation is approximately 2.07.

Standard Deviation Based on a Sample

When the given data is information obtained from a sample of the population, the formula for standard deviation is obtained by dividing the sum of the squares of the deviation from the mean by 1 less than the number of data values.

If each data value is listed separately:

$$\text{Standard deviation for a sample} = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$$

If the data is grouped in terms of the frequency of a given value:

$$\text{Standard deviation for a sample} = \sqrt{\frac{\sum_{i=1}^n f_i (x_i - \bar{x})^2}{\left(\sum_{i=1}^n f_i\right) - 1}}$$

The symbol for the standard deviation for a set of data that represents a sample is s . Many calculators use the symbol s_x .

EXAMPLE 4

From a high school, ten students are chosen at random to report their number of online friends. The data is as follows: 15, 13, 12, 10, 9, 7, 5, 4, 3, and 2.

Solution The total number of online friends for these 10 students is 80 or a mean of 8 online friends ($\bar{x} = 8$).

Online Friends (x_i)	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
15	7	49
13	5	25
12	4	16
10	2	4
9	1	1
7	-1	1
5	-3	9
4	-4	16
3	-5	25
2	-6	36
		$\sum_{i=1}^{10} (x_i - \bar{x})^2 = 182$

$$\text{Standard deviation} = \sqrt{\frac{1}{10-1} \sum_{i=1}^{10} (x_i - \bar{x})^2} = \sqrt{\frac{1}{9}(182)} \approx 4.5 \quad \text{Answer}$$

EXAMPLE 5

In each of the following, tell whether the population or sample standard deviation should be used.

- a. In a study of the land areas of the states of the United States, the area of each of the 50 states is used.
- b. In a study of the heights of high school students in a school of 1,200 students, the heights of 100 students chosen at random were recorded.
- c. In a study of the heights of high school students in the United States, the heights of 100 students from each of the 50 states were recorded.

Solution a. Use the population standard deviation since every state is included. *Answer*
b. Use the sample standard deviation since only a portion of the total school population was included in the study. *Answer*
c. Use the sample standard deviation since only a portion of the total school population was included in the study. *Answer*

EXAMPLE 6

A telephone survey conducted in Monroe County obtained information about the size of the households. Telephone numbers were selected at random until a sample of 130 responses were obtained. The frequency chart at the right shows the result of the survey.

No. of People per Household	1	2	3	4	5	6	7	8	9
Frequency	28	37	45	8	7	3	1	0	1

Solution The table below can be used to find the mean and the standard deviation.

x_i	f_i	$f_i x_i$	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	$f_i(x_i - \bar{x})^2$
1	28	28	-1.6	2.56	71.68
2	37	74	-0.6	0.36	13.32
3	45	135	0.4	0.16	7.20
4	8	32	1.4	1.96	15.68
5	7	35	2.4	5.76	40.32
6	3	18	3.4	11.56	34.68
7	1	7	4.4	19.36	19.36
8	0	0	5.4	29.16	0
9	1	9	6.4	40.96	40.96
	$\sum_{i=1}^9 f_i = 130$	$\sum_{i=1}^9 f_i x_i = 338$			$\sum_{i=1}^9 f_i(x_i - \bar{x})^2 = 243.20$

$$\text{Mean} = \frac{\sum_{i=1}^9 f_i x_i}{\sum_{i=1}^9 f_i} = \frac{338}{130} = 2.6$$

$$\text{Standard deviation} = \sqrt{\frac{\sum_{i=1}^9 f_i (x_i - \bar{x})^2}{\left(\sum_{i=1}^9 f_i\right) - 1}} = \sqrt{\frac{243.20}{130 - 1}} = \sqrt{\frac{243.20}{129}} = 1.373051826$$

Calculator Solution Enter the number of members of the households in L_1 and the frequency for each data value in L_2 .

ENTER: **STAT** **►** **ENTER** **2nd**
L1 **,** **2nd** **L2** **ENTER**

DISPLAY:

```
1-Var Stats
x̄=2.6
Σx=338
Σx²=1122
sx=1.373051826
σx=1.367160663
↓n=130
```

The standard deviation based on the data from a sample is $s_x = 1.373051826$ or approximately 1.37. **Answer**

Exercises

Writing About Mathematics

1. The sets of data for two different statistical studies are identical. The first set of data represents the data for all of the cases being studied and the second represents the data for a sample of the cases being studied. Which set of data has the larger standard deviation? Explain your answer.
2. Elaine said that the variance is the square of the standard deviation. Do you agree with Elaine? Explain why or why not.

Developing Skills

In 3–9, the given values represent data for a population. Find the variance and the standard deviation for each set of data.

3. 9, 9, 10, 11, 5, 10, 12, 9, 10, 12, 6, 11, 11, 11
4. 11, 6, 7, 13, 5, 8, 7, 10, 9, 11, 13, 12, 9, 16, 10
5. 20, 19, 20, 17, 18, 19, 42, 41, 41, 39, 39, 40
6. 20, 101, 48, -5, 63, 31, 20, 50, 16, 14, -45, 9

7.

x_i	f_i
30	1
35	7
40	10
45	9
50	11
55	8
60	6

8.

x_i	f_i
2	0
4	1
6	6
8	10
10	13
12	21
14	7
16	1

9.

x_i	f_i
20	21
25	10
30	1
35	2
40	2
45	4
50	5
55	3
60	7

In 10–16, the given values represent data for a sample. Find the variance and the standard deviation based on this sample.

10. 6, 4, 9, 11, 4, 3, 22, 3, 7, 10

11. 12.1, 33.3, 45.5, 60.1, 94.2, 22.2

12. 15, 10, 16, 19, 10, 19, 14, 17

13. 1, 3, 5, 22, 30, 45, 50, 55, 60, 70

14.

x_i	f_i
55	11
50	15
45	4
40	1
35	14
30	12
25	4

15.

x_i	f_i
33	3
34	1
35	4
36	6
37	5
38	11
39	6

16.

x_i	f_i
1	3
2	3
3	3
4	3
5	3
6	3
7	3

Applying Skills

17. To commute to the high school in which Mr. Fedora teaches, he can take either the Line A or the Line B train. Both train stations are the same distance from his house and both stations report that, on average, they run 10 minutes late from the scheduled arrival time. However, the standard deviation for Line A is 1 minute and the standard deviation for Line B is 5 minutes. To arrive at approximately the same time on a regular basis, which train line should Mr. Fedora use? Explain.

18. A hospital conducts a study to determine if nurses need extra staffing at night. A random sample of 25 nights was used. The number of calls to the nurses' station each night is shown in the stem-and-leaf diagram to the right.

a. Find the variance.

b. Find the standard deviation.

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

19. The ages of all of the students in a science class are shown in the table. Find the variance and the standard deviation.

Age	Frequency
18	1
17	2
16	9
15	9

20. The table shows the number of correct answers on a test consisting of 15 questions. The table represents correct answers for a sample of the students who took the test. Find the standard deviation based on this sample.

Correct Answers	6	7	8	9	10	11	12	13	14	15
Frequency	2	1	3	3	5	8	8	5	4	1

21. The table shows the number of robberies during a given month in 40 different towns of a state. Find the standard deviation based on this sample

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

22. Products often come with registration forms. One of the questions usually found on the registration form is household income. For a given product, the data below represents a random sample of the income (in thousands of dollars) reported on the registration form. Find the standard deviation based on this sample.

38 40 26 42 39 25 40 40 39 36
46 41 43 47 49 43 39 35 43 37

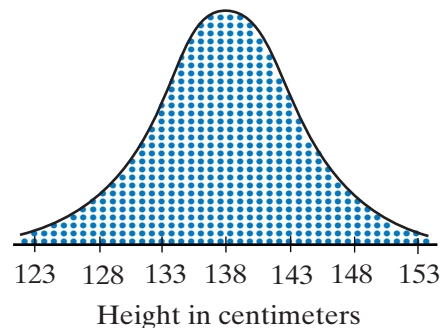
Hands-On Activity

The people in your survey from the Hands-On Activity of Section 15-1 represent a random sample of all people. Find the standard deviation based on your sample.

I 5-6 NORMAL DISTRIBUTION

The Normal Curve

Imagine that we were able to determine the height in centimeters of all 10-year-old children in the United States. With a scale along the horizontal axis that includes all of these heights, we will place a dot above each height for each child of that height. For example, we will place a dot above 140 on the horizontal scale a dot for each child who is 140 centimeters tall. Do this for 139, 138, 137, and so on for each height in our data. The result would be a type of frequency histogram. If we draw a smooth curve joining the top dot for each height, we will draw a bell-shaped curve called the **normal curve**. As the average height of 10-year-olds is approximately 138 centimeters, the data values are concentrated at 138 centimeters and the normal curve has a peak at 138 centimeters. Since for each height that is less than or greater than 138 centimeters there are fewer 10-year-olds, the normal curve progressively gets shorter as you go farther from the mean.

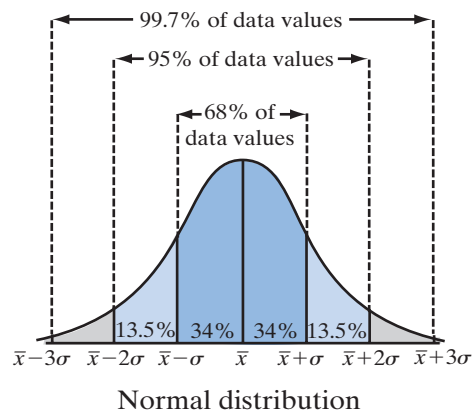


Scientists have found that large sets of data that occur naturally such as heights, weights, or shoe sizes have a bell-shaped or a normal curve. The highest point of the normal curve is at the mean of the data. The normal curve is symmetric with respect to a vertical line through the mean of the distribution.

Standard Deviation and the Normal Curve

A **normal distribution** is a set of data that can be represented by a normal curve. For a normal distribution, the following relationships exist.

1. The mean and the median of the data values lie on the line of symmetry of the curve.
2. Approximately 68% of the data values lie within one standard deviation from the mean.
3. Approximately 95% of the data values lie within two standard deviations from the mean.
4. Approximately 99.7% of the data values lie within three standard deviations from the mean.



EXAMPLE I

A set of data is normally distributed with a mean of 50 and a standard deviation of 2.

- What percent of the data values are less than 50?
- What percent of the data values are between 48 and 52?
- What percent of the data values are between 46 and 54?
- What percent of the data values are less than or equal to 46?

Solution a. In a normal distribution, 50% of the data values are to the left and 50% to the right of the mean.

- b. 48 and 52 are each 1 standard deviation away from the mean.

$$48 = 50 - 2 \quad 52 = 50 + 2$$

Therefore, 68% of the data values are between 48 and 52.

- c. 46 and 54 are each 2 standard deviations away from the mean.

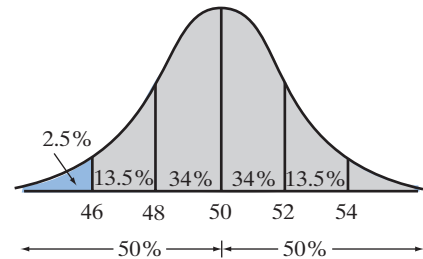
$$46 = 50 - 2(2) \quad 54 = 50 + 2(2)$$

Therefore, 95% of the data values are between 46 and 54.

- d. 50% of the data values are less than 50.

47.5% of the data values are more than 46 and less than 50.

Therefore, 50% - 47.5% or 2.5% of the data values are less than or equal to 46.



Answers a. 50% b. 68% c. 95% d. 2.5%

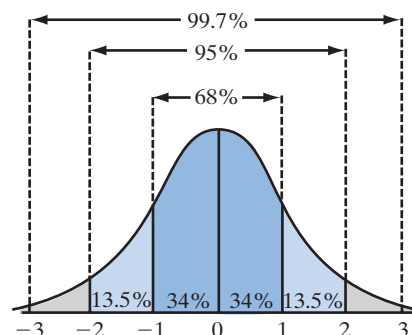
Z-Scores

The **z-score** for a data value is the deviation from the mean divided by the standard deviation. Let x be a data value of a normal distribution.

$$\text{z-score} = \frac{x - \bar{x}}{\text{standard deviation}} = \frac{x - \bar{x}}{\sigma}$$

The z-score of x , a value from a normal distribution, is positive when x is above the mean and negative when x is below the mean. The z-score tells us how many standard deviations x is above or below the mean.

1. The z-score of the mean is 0.
2. Of the data values, 34% have a z-score between 0 and 1 and 34% have a z-score between -1 and 0. Therefore, 68% have a z-score between -1 and 1.
3. Of the data values, 13.5% have a z-score between 1 and 2 and 13.5% have a z-score between -2 and -1.
4. Of the data values, $(34 + 13.5)\%$ or 47.5% have a z-score between 0 and 2 and 47.5% have a z-score between -2 and 0. Therefore, 95% have a z-score between -2 and 2.
5. Of the data values, 99.7% have a z-score between -3 and 3.



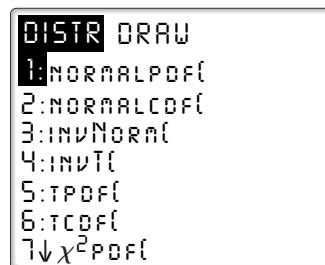
For example, the mean height of 10-year-old children is 138 centimeters with a standard deviation of 5. Casey is 143 centimeters tall.

$$\text{z-score for Casey} = \frac{143 - 138}{5} = \frac{5}{5} = 1$$

Casey's height is 1 standard deviation above the mean. For a normal distribution, 34% of the data is between the mean and 1 standard deviation above the mean and 50% of the data is below the mean. Therefore, Casey is as tall as or taller than $(34 + 50)\%$ or 84% of 10-year-old children.



A calculator will give us this same answer. The second entry of the DISTR menu is normalcdf(. When we use this function, we must supply a minimum value, a maximum value, the mean, and the standard deviation separated by commas:



$\text{normalcdf}(\text{minimum}, \text{maximum}, \text{mean}, \text{standard deviation})$

For the minimum value we can use 0. To find the proportion of 10-year-old children whose height is 143 centimeters or less, use the following entries.

ENTER: **2nd** **DISTR** **2** 0
 , 143 , 138 ,
 5 **)** **ENTER**

DISPLAY:

NORMALCDF(0,143,
 138,5)
 .8413447404

The calculator returns the number 0.8413447404, which can be rounded to 0.84 or 84%.

If we wanted to find the proportion of the 10-year-old children who are between 134.6 and 141.4 centimeters tall, we could make the following entry:

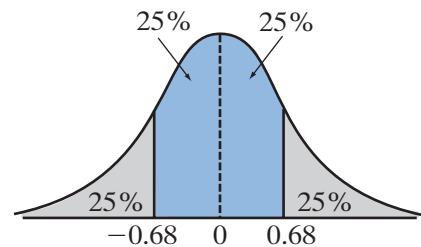
ENTER: 2nd DISTR 2 134.6
 , 141.4 , 138 ,
 5) ENTER

DISPLAY:

```
NORMALCDF(134.6,  
141.4,138,5)  
.5034956838
```

The calculator returns the number 0.5034956838, which can be rounded to 0.50 or 50%. Since 134.6 and 141.4 are equidistant from the mean, 25% of the data is below the mean and 25% is above the mean. Therefore, for this distribution, 134.6 centimeters is the first quartile, 138 centimeters is the median or second quartile, and 141.4 centimeters is the third quartile.

Note: 134.6 is 0.68 standard deviation below the mean and 141.4 is 0.68 standard deviation above the mean. For any normal distribution, data values with z-scores of -0.68 are approximately equal to the first quartile and data values with z-scores of 0.68 are approximately equal to the third quartile.



EXAMPLE 1

On a standardized test, the test scores are normally distributed with a mean of 60 and a standard deviation of 6.

- Of the data, 84% of the scores are at or below what score?
- Of the data, 16% of the scores are at or below what score?
- What is the z-score of a score of 48?
- If 2,000 students took the test, how many would be expected to score at or below 48?

Solution a. Since 50% scored at or below the mean and 34% scored within 1 standard deviation above the mean, $(50 + 34)\%$ or 84% scored at or below 1 deviation above the mean:

$$\bar{x} + \sigma = 60 + 6 = 66$$

- b.** Since 50% scored at or below the mean and 34% scored within 1 standard deviation below the mean, $(50 - 34)\%$ or 16% scored at or below 1 deviation below the mean:

$$\bar{x} + \sigma = 60 - 6 = 54$$

$$\text{c. z-score} = \frac{x - \bar{x}}{\sigma} = \frac{48 - 60}{6} = \frac{-12}{6} = -2$$

- d. A test score of 48 has a z-score of -2 . Since 47.5% of the scores are between -2 and 0 and 50% of the scores are less than 0, $50\% - 47.5\%$ or 2.5% scored at or below 48.

$$2.5\% \times 2,000 = 0.025 \times 2,000 = 50 \text{ students}$$

Answers a. 66 b. 54 c. -2 d. 50

EXAMPLE 2

For a normal distribution of weights, the mean weight is 160 pounds and a weight of 186 pounds has a z-score of 2.

- a. What is the standard deviation of the set of data?
b. What percent of the weights are between 155 and 165?

Solution a. z-score $= \frac{x - \bar{x}}{\sigma}$

$$2 = \frac{186 - 160}{\sigma}$$

$$2\sigma = 26$$

$$\sigma = 13 \text{ Answer}$$

b. ENTER: **2nd** **DISTR** **2**
155 **,** 165 **,** 160 **,**
13 **)** **ENTER**

DISPLAY: `NORMALCDF(155,165,160,13)`
.2994775047

About 30% of the weights are between 155 and 165. Answer

Exercises

Writing About Mathematics

1. A student's scores on five tests were 98, 97, 95, 93, and 67. Explain why this set of scores does not represent a normal distribution.
2. If 34% of the data for a normal distribution lies between the mean and 1 standard deviation above the mean, does 17% of the data lie between the mean and one-half standard deviation above the mean? Justify your answer.

Developing Skills

In 3–9, for a normal distribution, determine what percent of the data values are in each given range.

3. Between 1 standard deviation below the mean and 1 standard deviation above the mean
4. Between 1 standard deviation below the mean and 2 standard deviations above the mean

5. Between 2 standard deviations below the mean and 1 standard deviation above the mean
6. Above 1 standard deviation below the mean
7. Below 1 standard deviation above the mean
8. Above the mean
9. Below the mean
10. A set of data is normally distributed with a mean of 40 and a standard deviation of 5. Find a data value that is:
 - a. 1 standard deviation above the mean
 - b. 2.4 standard deviations above the mean
 - c. 1 standard deviation below the mean
 - d. 2.4 standard deviations below the mean

Applying Skills

In 11–14, select the numeral that precedes the choice that best completes the statement or answers the question.

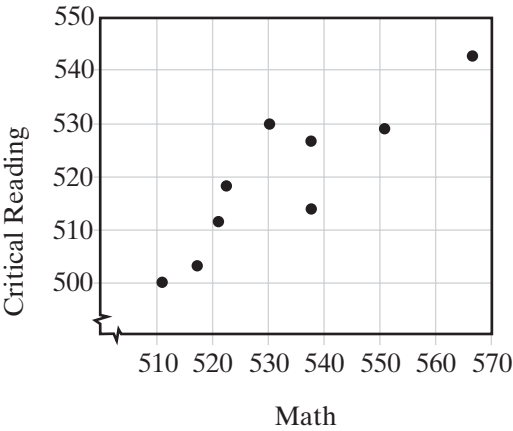
11. The playing life of a Euclid mp3 player is normally distributed with a mean of 30,000 hours and a standard deviation of 500 hours. Matt's mp3 player lasted for 31,500 hours. His mp3 player lasted longer than what percent of other Euclid mp3 players?
(1) 68% (2) 95% (3) 99.7% (4) more than 99.8%
12. The scores of a test are normally distributed. If the mean is 50 and the standard deviation is 8, then a student who scored 38 had a z-score of
(1) 1.5 (2) -1.5 (3) 12 (4) -12
13. The heights of 10-year-old children are normally distributed with a mean of 138 centimeters with a standard deviation of 5 centimeters. The height of a 10-year-old child who is as tall as or taller than 95.6% of all 10-year-old children is
(1) between 138 and 140 cm. (2) between 140 and 145 cm.
(3) between 145 and 148 cm. (4) taller than 148 cm.
14. The heights of 200 women are normally distributed. The mean height is 170 centimeters with a standard deviation of 10 centimeters. What is the best estimate of the number of women in this group who are between 160 and 170 centimeters tall?
(1) 20 (2) 34 (3) 68 (4) 136
15. When coffee is packed by machine into 16-ounce cans, the amount can vary. The mean weight is 16.1 ounces and the standard deviation is 0.04 ounce. The weight of the coffee approximates a normal distribution.
 - a. What percent of the cans of coffee can be expected to contain less than 16 ounces of coffee?
 - b. What percent of the cans of coffee can be expected to contain between 16.0 and 16.2 ounces of coffee?

16. The length of time that it takes Ken to drive to work represents a normal distribution with a mean of 25 minutes and a standard deviation of 4.5 minutes. If Ken allows 35 minutes to get to work, what percent of the time can he expect to be late?
17. A librarian estimates that the average number of books checked out by a library patron is 4 with a standard deviation of 2 books. If the number of books checked out each day approximates a normal distribution, what percent of the library patrons checked out more than 7 books yesterday?
18. The heights of a group of women are normally distributed with a mean of 170 centimeters and a standard deviation of 10 centimeters. What is the z-score of a member of the group who is 165 centimeters tall?
19. The test grades for a standardized test are normally distributed with a mean of 50. A grade of 60 represents a z-score of 1.25. What is the standard deviation of the data?
20. Nora scored 88 on a math test that had a mean of 80 and a standard deviation of 5. She also scored 80 on a science test that had a mean of 70 and a standard deviation of 3. On which test did Nora perform better compared with other students who took the tests?

I 5-7 BIVARIATE STATISTICS

Statistics are often used to compare two sets of data. For example, a pediatrician may compare the height and weight of a child in order to monitor growth. Or the owner of a gift shop may record the number of people who enter the store with the revenue each day. Each of these sets of data is a pair of numbers and is an example of **bivariate statistics**.

Representing bivariate statistics on a two-dimensional graph or **scatter plot** can help us to observe the relationship between the variables. For example the mean value for the critical reading and for the math sections of the SAT examination for nine schools in Ontario County are listed in the table and shown on the graph on the right.



Math	530	551	521	522	537	511	516	537	566
Critical Reading	530	529	512	518	526	500	504	515	543

The graph shows that there appears to be a linear relationship between the critical reading scores and the math scores. As the math scores increase, the critical reading scores also increase. We say that there is a **correlation** between the two scores. The points of the graph approximate a line.



These data can also be shown on a calculator. Enter the math scores as L_1 and the corresponding critical reading scores as L_2 . Then turn on Plot 1 and use ZoomStat from the ZOOM menu to construct a window that will include all values of x and y :

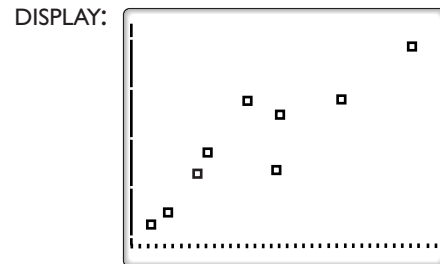
ENTER: **2nd** **STAT PLOT** **1**
ENTER **▼** **ENTER**
▼ **2nd** **L1** **▼**
2nd **L2**

ENTER: **ZOOM** **9**

DISPLAY:

```

PLOT1 PLOT2 PLOT3
ON  OFF
TYPE: [scatter] [line] [histogram]
XLIST:L1
YLIST:L2
MARK: [square] +
    
```



The calculator will display a graph similar to that shown above. To draw a line that approximates the data, use the **regression line** on the calculator. A regression line is a special **line of best fit** that minimizes the square of the vertical distances to each data point. In this course, you do not have to know the formula to find the regression line. The calculator can be used to determine the regression line:

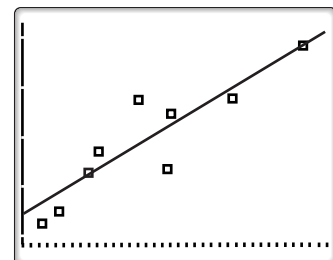
```

LINREG
Y=AX+B
A=.6925241158
B=151.0129957
R^2=.7986740639
R=.8936856628
    
```

ENTER: **STAT** **►** **4** **VARS** **►** **1** **1** **ENTER**

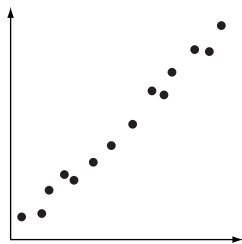
The calculator displays values for a and b for the linear equation $y = ax + b$ and stores the regression equation into Y_1 in the **Y=** menu. Press **ZOOM** **9** to display the scatter plot and the line that best approximates the data. If we round the given values of a and b to three decimal places, the linear regression equation is:

$$y = 0.693x + 151.013$$

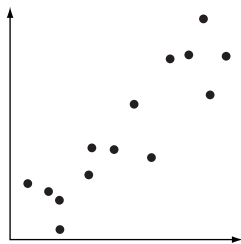


We will study other data that can be approximated by a curve rather than a line in later sections.

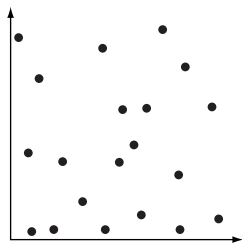
The scatter plots below show possible linear correlation between elements of the pairs of bivariate data. The correlation is positive when the values of the second element of the pairs tend to increase when the values of the first elements of the pairs increase. The correlation is negative when the values of the second element of the pairs tend to decrease when the values of the first elements of the pairs increase.



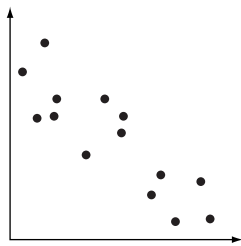
Strong positive correlation



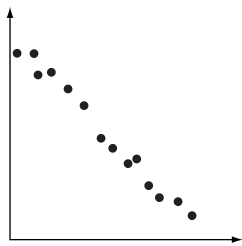
Moderate positive correlation



No linear correlation



Moderate negative correlation



Strong negative correlation

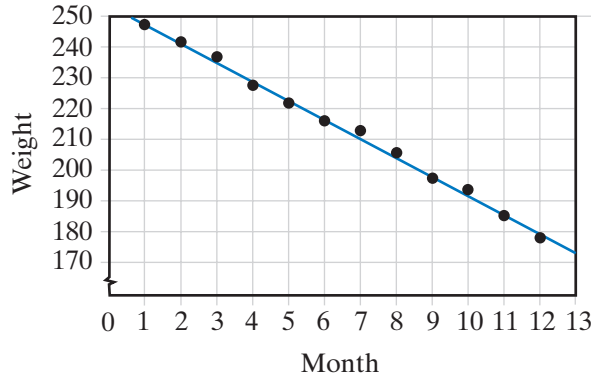
EXAMPLE I

Jacob joined an exercise program to try to lose weight. Each month he records the number of months in the program and his weight at the end of that month. His record for the first twelve months is shown below:

Month	1	2	3	4	5	6	7	8	9	10	11	12
Weight	248	242	237	228	222	216	213	206	197	193	185	178

- Draw a scatter plot and describe the correlation between the data (if any).
- Draw a line that appears to represent the data.
- Write an equation of a line that best represents the data.

Solution a. Use a scale of 0 to 13 along a horizontal line for the number of months in the program and a scale from 170 to 250 along the vertical axis for his weight at the end of that month. The scatter plot is shown here. There appears to be a strong negative correlation between the number of months and Jacob's weight.



- The line on the graph that appears to approximate the data intersects the points (1, 248) and (11, 185). We can use these two points to write an equation of the line.

$$y - 248 = \frac{248 - 185}{1 - 11}(x - 1)$$

$$y - 248 = -6.3x + 6.3$$

$$y = -6.3x + 254.3$$

- On a calculator, enter the number of the month in L_1 and the weight in L_2 . To use the calculator to determine a line that approximates the data, use the following sequence:

ENTER: **STAT** **►** **4** **VARS** **►** **1**
1 **ENTER**

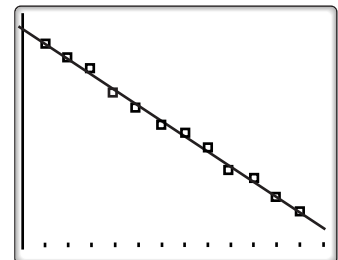
```
LINREG
Y=AX+B
A=-6.283216783
B=254.5909091
R^2=.9966845178
R=.9983408825
```

The calculator displays values for a and b for the linear equation $y = ax + b$ and stores the equation as Y_1 in the **Y=** menu. If we round the given values of a and b to three decimal places, the linear regression equation is:

$$y = -6.283x + 254.591$$

Turn on Plot 1, then use ZoomStat to graph the data:

ENTER: **2nd** **STAT PLOT** **ENTER** **ENTER**
▼ **ENTER** **▼** **2nd** **L1**
ENTER **2nd** **L2** **ENTER**
ZOOM **9**



The calculator will display the scatter plot of the data along with the regression equation.

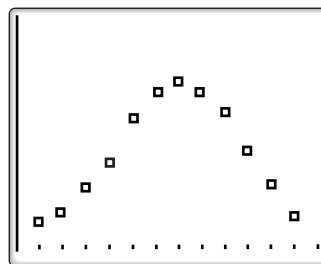
EXAMPLE 2

In order to assist travelers in planning a trip, a travel guide lists the average high temperature of most major cities. The listing for Albany, New York is given in the following table.

Month	1	2	3	4	5	6	7	8	9	10	11	12
Temperature	31	33	42	51	70	79	84	81	73	62	48	35

Can these data be represented by a regression line?

Solution Draw a scatter plot the data. The graph shows that the data can be characterized by a curve rather than a line. Finding the regression line for this data would not be appropriate.

**SUMMARY**

- The slope of the regression line gives the direction of the correlation:
 - A positive slope shows a positive correlation.
 - A negative slope shows a negative correlation.
- The regression line is appropriate only for data that appears to be linearly related. Do *not* calculate a regression line for data with a scatter plot showing a non-linear relationship.
- The regression equation is sensitive to rounding. Round the coefficients to at least three decimal places.

Exercises**Writing About Mathematics**

1. Explain the difference between univariate and bivariate data and give an example of each.
2. What is the relationship between slope and correlation? Can slope be used to measure the strength of a correlation? Explain.

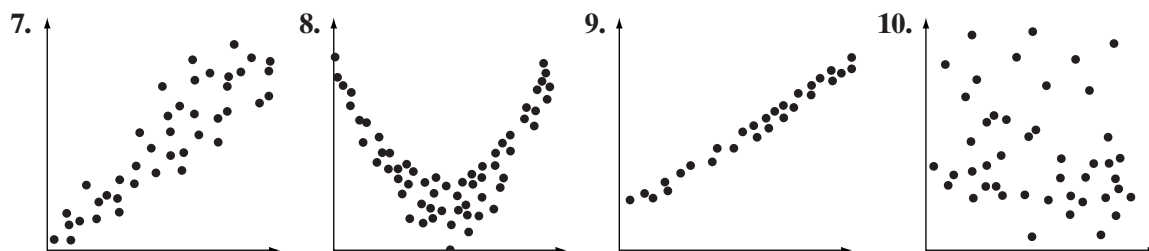
Developing Skills

In 3–6, is the set of data to be collected univariate or bivariate?

3. The science and math grades of all students in a school
4. The weights of the 56 first-grade students in a school

5. The weights and heights of the 56 first-grade students in a school
6. The number of siblings for each student in the first grade

In 7–10, look at the scatter plots and determine if the data sets have high linear correlation, moderate linear correlation, or no linear correlation.



Applying Skills

In 11–17: **a.** Draw a scatter plot. **b.** Does the data set show strong positive linear correlation, moderate positive linear correlation, no linear correlation, moderate negative linear correlation, or strong negative linear correlation? **c.** If there is strong or moderate correlation, write the equation of the regression line that approximates the data.

11. The following table shows the number of gallons of gasoline needed to fill the tank of a car and the number of miles driven since the previous time the tank was filled.

Gallons	8.5	7.6	9.4	8.3	10.5	8.7	9.6	4.3	6.1	7.8
Miles	255	230	295	250	315	260	290	130	180	235

12. A business manager conducted a study to examine the relationship between number of ads placed for each month and sales. The results are shown below where sales are in the thousands.

Number of Ads	10	12	14	16	18	20	22	24	26	28	30
Sales	20	26.5	32	34.8	40	47.2	49.1	56.9	57.9	65.8	66.4

13. Jack Sheehan looked through some of his favorite recipes to compare the number of calories per serving to the number of grams of fat. The table below shows the results.

Calories	310	210	260	330	290	320	245	293	220	260	350
Fat	11	5	11	12	14	16	7	10	8	8	15

14. Greg did a survey to support his theory that the size of a family is related to the size of the family in which the mother of the family grew up. He asked 20 randomly selected people to list the number of their siblings and the number of their mother's siblings. Greg made the following table.

Family	2	2	3	1	0	3	5	2	1	2
Mother's Family	4	0	3	7	4	2	2	6	4	7
Family	3	6	0	2	1	4	1	0	4	1
Mother's Family	5	4	6	1	0	2	1	3	3	2

15. When Marie bakes, it takes about five and a half minutes for the temperature of the oven to reach 350°. One day, while waiting for the oven to heat, Marie recorded the temperature every 20 seconds. Her record is shown below.

Seconds	0	20	40	60	80	100	120	140	160
Temperature	100	114	126	145	160	174	193	207	222
Seconds	180	200	220	240	260	280	300	320	340
Temperature	240	255	268	287	301	318	331	342	350

16. An insurance agent is studying the records of his insurance company looking for a relationship between age of a driver and the percentage of accidents due to speeding. The table shown below summarizes the findings of the insurance agent.

Age	17	18	21	25	30	35	40	45	50	55	60	65
% of Speeding Accidents	49	49	48	39	31	33	24	25	16	10	5	6

17. A sociologist is interested in the relationship between body weight and performance on the SAT. A random sample of 10 high school students from across the country provided the following information:

Weight	197	193	194	157	159	170	149	169	157	185
Score	1,485	1,061	1,564	1,729	1,668	1,405	1,544	1,752	1,395	1,214

15-8 CORRELATION COEFFICIENT

We would like to measure the strength of the linear relationship between the variables in a set of bivariate data. The slope of the regression equation tells us the direction of the relationship but it does not tell us the strength of the relationship. The number that we use to measure both the strength and direction of the linear relationship is called the **correlation coefficient**, r . The value of the correlation coefficient does not depend on the units of measurement. In more advanced statistics courses, you will learn a formula to derive the correlation coefficient. In this course, we can use the graphing calculator to calculate the value of r .

EXAMPLE 1

The coach of the basketball team made the following table of attempted and successful baskets for eight players.

Attempted Baskets (x_i)	10	12	12	13	14	15	17	19
Successful Baskets (y_i)	6	7	9	8	10	11	14	15

Find the value of the correlation coefficient.

Solution Enter the given x_i in L_1 and y_i in L_2 .

Then choose LinReg(ax+b) from the CALC STAT menu:

ENTER: **STAT** **►** **4** **ENTER**

The calculator will list both the regression equation and r , the correlation coefficient.

```

LINREG
Y=AX+B
A=1.066666667
B=-4.933333333
R2=.9481481481
R=.9737289911

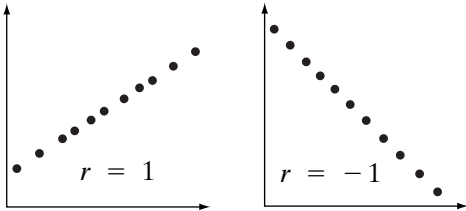
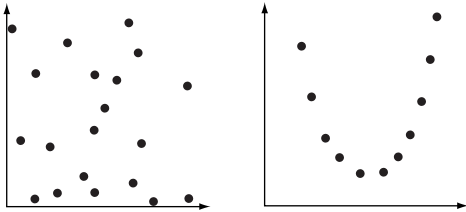
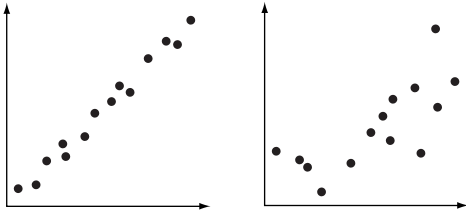
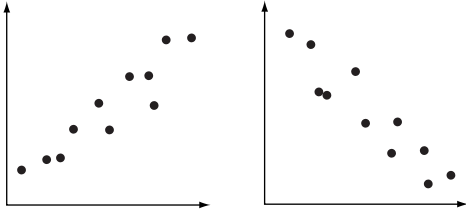
```

Answer $r = 0.97$



Note: If the correlation coefficient does not appear on your calculator, enter **2nd** **CATALOG** **D**, scroll down to DiagnosticOn, press **ENTER**, and press **ENTER** again.

When the absolute value of the correlation coefficient is close to 1, the data have a strong linear correlation. When the absolute value of the correlation coefficient is close to 0, there is little or no linear correlation. Values between 0 and 1 indicate various degrees of positive moderate correlation and values between 0 and -1 indicate various degrees of negative moderate correlation.

Properties of the Correlation Coefficient, r	
$-1 \leq r \leq 1$. The correlation coefficient is a number between -1 and 1 .	
When $r = 1$ or -1 , there is a perfect linear relationship between the data values.	
When $r = 0$, no linear relationship exists between the data values.	
When $ r $ is close to 1 , the data have a strong linear relationship. Values between 0 and 1 indicate various degrees of moderate correlation.	
The sign of r matches the sign of the slope of the regression line.	

EXAMPLE 2

An automotive engineer is studying the fuel efficiency of a new prototype. From a fleet of eight prototypes, he records the number of miles driven and the number of gallons of gasoline used for each trip.

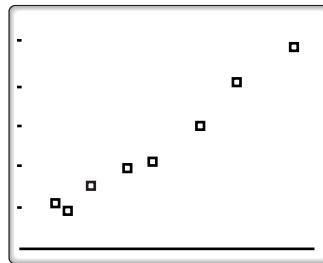
Miles Driven	310	270	350	275	380	320	290	405
Gallons of Gasoline Used	10.0	9.0	11.2	8.7	12.3	10.2	9.5	12.7

- Based on the context of the problem, do you think the correlation coefficient will be positive, negative, or close to 0?
- Based on the scatter plot of the data, do you expect the correlation coefficient to be close to -1 , 0 , or 1 ?
- Use a calculator to find the equation of the regression line and determine the correlation coefficient.

Solution a. As gallons of gasoline used tend to increase with miles driven, we expect the correlation coefficient to be positive.

Enter the given data as L_1 and L_2 on a calculator.

b.



There appears to be a strong positive correlation, so r will be close to 1.

c.

```
LINREG
Y=AX+B
A=.0298533724
B=.7476539589
R^2=.9879952255
R=.9939794895
```

$$y = 0.030x + 0.748, r = 0.99$$

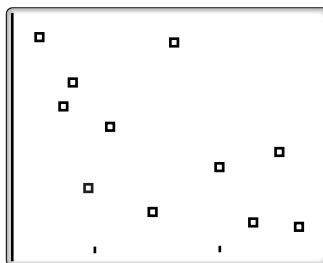
EXAMPLE 3

The produce manager of a food store noted the relationship between the amount the store charged for a pound of fresh broccoli and the number of pounds sold in one week. His record for 11 weeks is shown in the following table.

Cost per Pound	\$0.65	\$0.85	\$0.90	\$1.00	\$1.25	\$1.50	\$1.75	\$1.99	\$2.25	\$2.50	\$2.65
Pounds Purchased	58	43	49	23	39	16	56	32	12	35	11

What conclusion could the product manager draw from this information?

Solution Enter the given data as L_1 and L_2 on the calculator, graph the scatter plot, and find the value of the correlation coefficient.



```

LINREG
Y=AX+B
A=-13.82881393
B=55.73638117
R^2=.331571737
R=-.5758226611

```

From the scatter plot, we can see that there is a moderate negative correlation. Since $r = -0.58$, the product manager might conclude that a lower price does explain some of the increase in sales but other factors also influence the number of sales.

A Warning About Cause-and-Effect

The correlation coefficient is a number that measures the strength of the linear relationship between two data sets. However, simply because there appears to be a strong linear correlation between two variables does *not* mean that one causes the other. There may be other variables that are the cause of the observed pattern. For example, consider a study on the population growth of a city. Although a statistician may find a linear pattern over time, this does not mean that time causes the population to grow. Other factors cause the city grow, for example, a booming economy.

SUMMARY

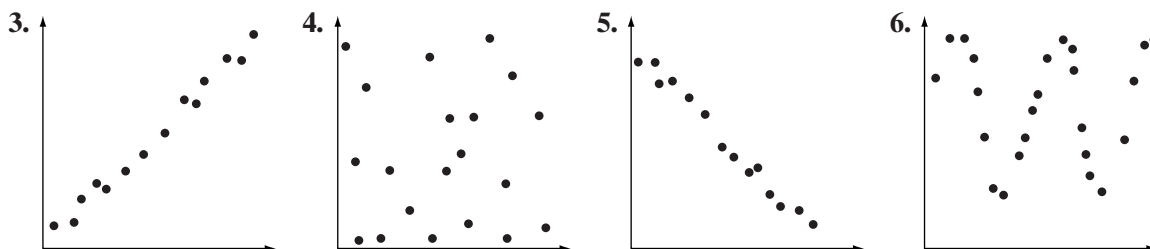
- $-1 \leq r \leq 1$. The correlation coefficient is a number between -1 and 1 .
- When $r = 1$ or $r = -1$, there is a perfect linear relationship between the data values.
- When $r = 0$, no *linear* relationship exists between the data values.
- When $|r|$ is close to 1 , the data have a strong linear relationship. Values between 0 and 1 indicate various degrees of moderate correlation.
- The sign of r matches the sign of the slope of the regression line.
- A high correlation coefficient does *not* necessarily mean that one variable causes the other.

Exercises**Writing About Mathematics**

1. Does a correlation coefficient of -1 indicate a lower degree of correlation than a correlation coefficient of 0 ? Explain why or why not.
2. If you keep a record of the temperature in degrees Fahrenheit and in degrees Celsius for a month, what would you expect the correlation coefficient to be? Justify your answer.

Developing Skills

In 3–6, for each of the given scatter plots, determine whether the correlation coefficient would be close to -1 , 0 , or 1 .



In 7–14, for each of the given correlation coefficients, describe the linear correlation as strong positive, moderate positive, none, moderate negative, or strong negative.

- | | | | |
|--------------|----------------|---------------|-----------------|
| 7. $r = 0.9$ | 8. $r = -1$ | 9. $r = -0.1$ | 10. $r = 0.3$ |
| 11. $r = 1$ | 12. $r = -0.5$ | 13. $r = 0$ | 14. $r = -0.95$ |

Applying Skills

In 15–19: **a.** Draw a scatter plot for each data set. **b.** Based on the scatter plot, would the correlation coefficient be close to -1 , 0 , or 1 ? Explain. **c.** Use a calculator to find the correlation coefficient for each set of data.

15. The following table shows the number of gallons of gasoline needed to fill the tank of a car and the number of miles driven since the previous time the tank was filled.

Gallons	12.5	3.4	7.9	9.0	15.7	7.0	5.1	11.9	13.0	10.7
Miles	392	137	249	308	504	204	182	377	407	304

16. A man on a weight-loss program tracks the number of pounds that he lost over the course of 10 months. A negative number indicates that he actually *gained* weight for that month.

Month	1	2	3	4	5	6	7	8	9	10
No. of Pounds Lost	9.2	9.1	4.8	4.5	2.8	1.8	1.2	0	0.8	-2.6

17. An economist is studying the job market in a large city conducts of survey on the number of jobs in a given neighborhood and the number of jobs paying \$100,000 or more a year. A sample of 10 randomly selected neighborhood yields the following data:

Total Number of Jobs	24	28	17	39	32	21	39	39	24	29
No. of High-Paying Jobs	3	3	4	5	7	3	4	7	7	4

18. The table below shows the same-day forecast and the actual high temperature for the day over the course of 18 days. The temperature is given in degrees Fahrenheit.

Same-Day Forecast	56	52	67	55	58	56	59	57	53
Actual Temperature	53	54	63	49	66	54	54	56	59
Same-Day Forecast	45	55	45	58	59	55	48	53	54
Actual Temperature	48	60	36	59	59	47	46	52	48

19. The table below shows the five-day forecast and the actual high temperature for the fifth day over the course of 18 days. The temperature is given in degrees Fahrenheit.

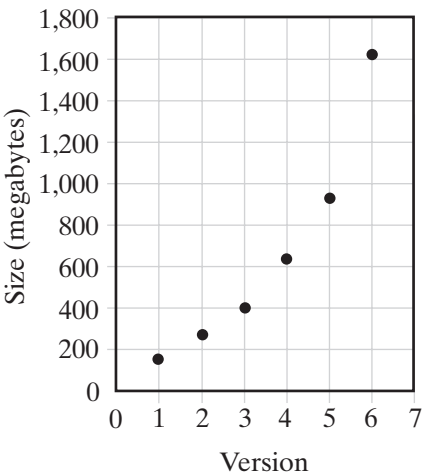
Five-Day Forecast	56	52	67	55	58	56	59	57	53
Actual Temperature	50	50	84	54	57	40	70	79	48
Five-Day Forecast	45	55	45	58	59	55	48	53	54
Actual Temperature	40	61	40	70	46	75	49	46	88

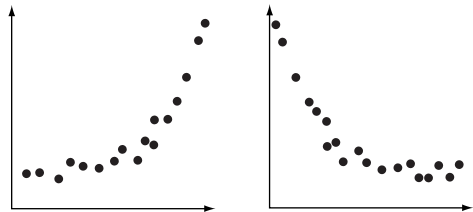
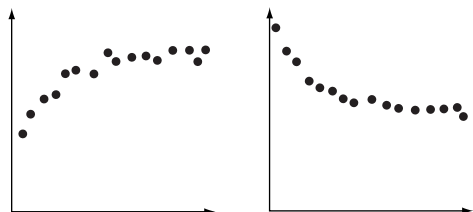
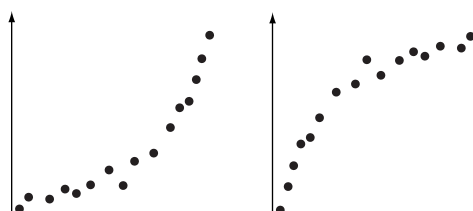
20. a. In Exercises 18 and 19, if the forecasts were 100% accurate, what should the value of r be?
- b. Is the value of r for Exercise 18 greater than, equal to, or less than the value of r for Exercise 19? Is this what you would expect? Explain.

I 5-9 NON-LINEAR REGRESSION

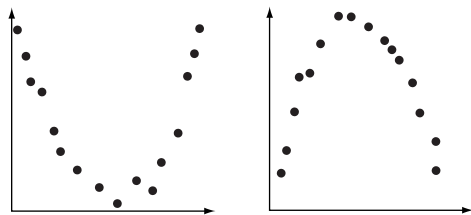
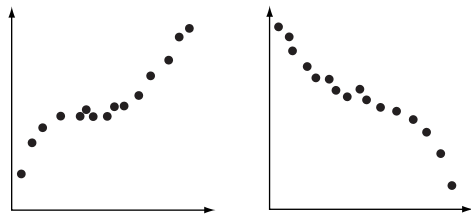
Not all bivariate data can be represented by a linear function. Some data can be better approximated by a curve. For example, on the right is a scatter plot of the file size of a computer program called Super Type over the course of 6 different versions. The relationship does not appear to be linear. For this set of data, a linear regression would *not* be appropriate.

There are a variety of non-linear functions that can be applied to non-linear data. In a statistics course, you will learn more rigorous methods of determining the regression model. In this course, we will use the scatter plot of the data to choose the regression model:



Regression to Use	Description of Scatter Plot	Examples
Exponential	<ul style="list-style-type: none">• An exponential curve that does <i>not</i> pass through (0, 0)• y-intercept is positive• Data constraint: $y > 0$	
Logarithmic	<ul style="list-style-type: none">• A logarithmic curve that does <i>not</i> pass through (0, 0)• y-intercept is positive or negative• Data constraint: $x > 0$	
Power	<ul style="list-style-type: none">• Positive half of power curve passing through (0, 0)• Data constraints: $x > 0, y > 0$	

(continued on next page)

Regression to Use	Description of Scatter Plot	Examples
<i>Specific types of power regression:</i>		
Quadratic	• A quadratic curve	
Cubic	• A cubic curve	



The different non-linear regression models can be found in the STAT CALC menu of the graphing calculator.

- 5:QuadReg is quadratic regression.
- 6:CubicReg is cubic regression.
- 9:LnReg is logarithmic regression.
- 0:ExpReg is exponential regression.
- A:PwrReg is power regression.

In the example of the file size of Super Type, the scatter plot appears to be exponential or power. The table below shows the data of the scatter plot:

Version	1	2	3	4	5	6
Size (megabytes)	155	240	387	630	960	1,612

To find the exponential regression model, enter the data into L_1 and L_2 . Choose ExpReg from the STAT CALC menu:

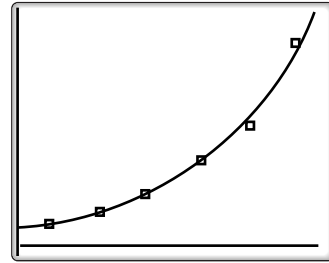
ENTER: **STAT** **►** **0** **VARS** **►** **1** **1** **ENTER**

The calculator will display the regression equation and store the equation into Y_1 of the **Y=** menu. To the nearest thousandth, the regression equation is $y = 95.699(1.596^x)$. Press **ZOOM** **9** to graph the scatter plot and the regression equation.

```

EXPREG
Y=A*B^X
A=95.69902108
B=1.595666689
R^2=.9984598905
R=.9997299088

```



To find the power regression model, with the data in L_1 and L_2 , choose PwrReg from the STAT CALC menu:

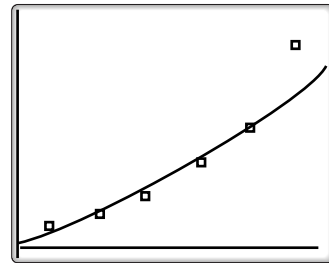
ENTER: **STAT** **▶** **ALPHA** **A** **VARS** **▶** **1** **1** **ENTER**

The calculator will display the regression equation and store the equation into Y_1 of the **Y=** menu. To the nearest thousandth, the regression equation is $y = 121.591x^{1.273}$. Press **ZOOM** **9** to graph the scatter plot and the regression equation.

```

PWRREG
Y=A*X^B
A=121.5914377
B=1.273149963
R^2=.9307655743
R=.9647619263

```



From the scatter plots, we see that the exponential regression equation is a better fit for the data.

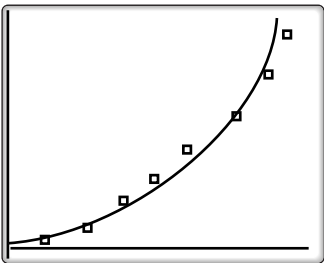
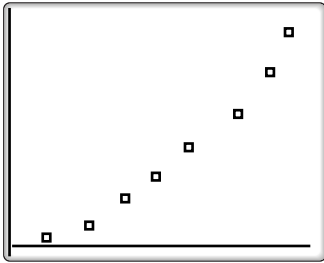
EXAMPLE I

A stone is dropped from a height of 1,000 feet. The trajectory of the stone is recorded by a high-speed video camera in intervals of half a second. The recorded distance that the stone has fallen in the first 5 seconds is given below:

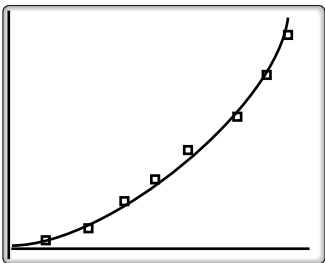
Seconds	1	1.5	2	2.5	3	3.5	4	4.5
Distance	16	23	63	105	149	191	260	321

- Determine which regression model is most appropriate.
- Find the regression equation. Round all values to the nearest thousandth.

Solution a. Draw a scatter plot of the data. The data appears to approximate an exponential function or a power function. Enter the data in L_1 and L_2 . Find and graph the exponential and power models. From the displays, it appears that the power model is the better fit.



exponential



power

b. To the nearest thousandth, the calculator will display the power equation $y = ax^b$ for $a = 13.619$, $b = 2.122$.

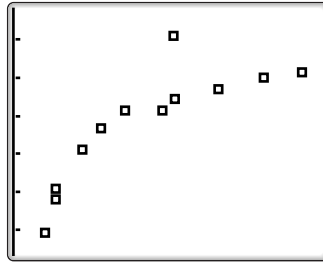
Answers a. Power regression **b.** $y = 13.619(x^{2.122})$

EXAMPLE 2

A pediatrician has the following table that lists the head circumferences for a group of 12 baby girls from the same extended family. The circumference is given in centimeters.

Age in Months	2	2	5	4	1	17	11	14	7	11	10	19
Circumference	36.8	37.2	38.6	38.2	35.9	40.4	39.7	39.9	39.2	41.1	39.3	40.5

- a.** Make a scatter plot of the data.
- b.** Choose what appears to be the curve that best fits the data.
- c.** Find the regression equation for this model.

Solution a.**b.** The data appears to approximate a log function. *Answer***c.** With the ages in L_1 and the circumferences in L_2 , choose LnReg from the STAT CALC menu:ENTER: **STAT** **►** **9** **ENTER**

The equation of the regression equation, to the nearest thousandth, is

$$y = 35.938 + 1.627 \ln x \quad \text{Answer}$$

```

LNREG
Y=A+BLNX
A=35.9381563
B=1.627161495
R^2=.9281619959
R=.9634116441

```

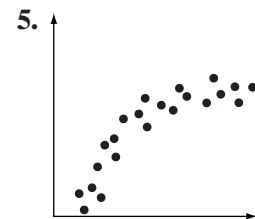
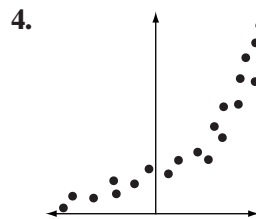
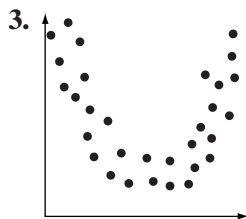
Exercises

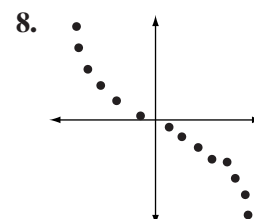
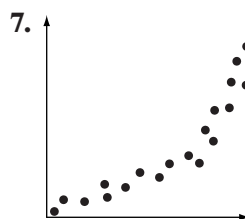
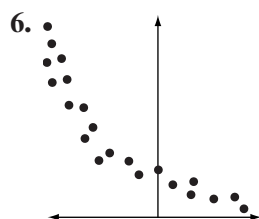
Writing About Mathematics

- At birth, the average circumference of a child's head is 35 centimeters. If the pair $(0, 35)$ is added to the data in Example 2, the calculator returns an error message. Explain why.
- Explain when the power function, $y = ax^b$, has only positive or only negative y -values and when it has both positive and negative y -values.

Developing Skills

In 3–8, determine the regression model that appears to be appropriate for the data.





In 9–13: **a.** Create a scatter plot for the data. **b.** Determine which regression model is the most appropriate for the data. Justify your answer. **c.** Find the regression equation. Round the coefficient of the regression equation to three decimal places.

9.

x	4	7	3	8	6	5	6	3	9	4.5
y	10	7	15	9	5	6	6	14	14	8

10.

x	-2.3	1.8	-1.0	4.6	1.4	3.7	5.3	1.9	0.7	4.2
y	2.2	18.3	8.2	63.7	15.3	43.0	89.5	22.7	12.1	54.7

11.

x	3.3	-3.8	-2.1	0.4	3.5	-3.8	-1.8	-0.4	2.4	1.2
y	12.5	-17.1	-3.6	0.4	15.0	-18.9	-2.1	-0.4	4.2	3.4

12.

x	1	2	3	4	5	6	7	8	9	10
y	-7	-5.6	-4.8	-4.2	-3.8	-3.4	-3.1	-2.8	-2.6	-2.4

13.

x	1	2	3	4	5	6	7	8	9	10
y	2.7	2.3	2.0	1.7	1.5	1.2	1.1	0.9	0.8	0.7

Applying Skills

14. Mrs. Vroman bought \$1,000 worth of shares in the Acme Growth Company. The table below shows the value of the investment over 10 years.

Year	1	2	3	4	5	6	7	8	9	10
Value (\$)	1,045	1,092	1,141	1,192	1,246	1,302	1,361	1,422	1,486	1,553

- Find the exponential regression equation for the data with the coefficient and base rounded to three decimal places.
- Predict, to the nearest dollar, the value of the Vromans' investment after 11 years.

15. The growth chart below shows the average height in inches of a group of 100 children from 2 months to 36 months.

Month	2	4	6	8	10	12	14	16	18
Height in Inches	22.7	26.1	27.5	28.9	32.1	31.7	33.1	32.7	34.0
Month	20	22	24	26	28	30	32	34	36
Height in inches	34.4	34.6	36.0	34.6	35.2	36.6	35.6	37.2	37.6

- Find the logarithmic regression equation for the data with the coefficients rounded to three decimal places.
 - Predict, to the nearest tenth of an inch, the average height of a child at 38 months.
16. The orbital speed in kilometers per second and the distance from the sun in millions of kilometers of each of six planets is given in the table.

Planet	Venus	Earth	Mars	Jupiter	Saturn	Uranus
Orbital Speed	34.8	29.6	23.9	12.9	9.6	6.6
Distance from the Sun	108.2	149.6	227.9	778.0	1,427	2,871

- Find the regression equation that appears to be the best fit for the data with the coefficient rounded to three decimal places.
 - Neptune has an orbital speed of 5.45 km/sec and is 4,504 million kilometers from the sun. Does the equation found for the six planets given in the table fit the data for Neptune?
17. A mail order company has shipping boxes that have square bases and varying heights from 1 to 5 feet. The relationship between the height of the box and the volume is shown in the table.

Height (ft)	1	1.5	2	2.5	3	3.5	4	4.5	5
Volume (ft ³)	2	7	16	31	54	86	128	182	250

- Create a scatter plot for the data. Let the horizontal axis represent the height of the box and the vertical axis represent the volume.
- Determine which regression model is most appropriate for the data. Justify your answer.
- Find the regression equation. Round the coefficient of the regression equation to three decimal places.

18. In an office building the thermostats have six settings. The table below shows the average temperature in degrees Fahrenheit for a month that each setting produced.

Setting	1	2	3	4	5	6
Temperature (°F)	61	64	66	67	69	70

- Create a scatter plot for the data. Let the horizontal axis represent the setting and the vertical axis represent temperature.
 - Find the equation of best fit using a power regression. Round the coefficient of the regression equation to three decimal places.
19. The following table shows the speed in megahertz of Intel computer chips over the course of 36 years. The time is given as the number of years since 1971.

Year	0	1	3	7	11	14	18	22
Speed	0.108	0.8	2	5	6	16	25	66
Year	24	26	28	29	31	34	35	36
Speed	200	300	500	1,500	1,700	3,200	2,900	3,000

One application of *Moore's Law* is that the speed of a computer processor should double approximately every two years. Use this information to determine the regression model. Does Moore's Law hold for Intel computer chips? Explain.

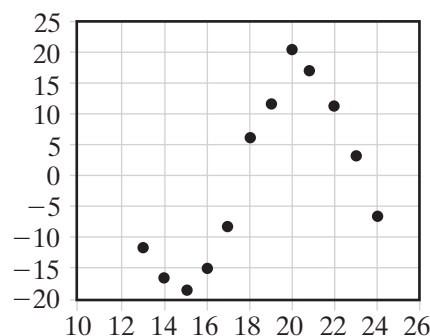
Hands-On Activity: Sine Regression

If we make a scatter plot of the following set of data on a graphing calculator, we may observe that the data points appear to form a sine curve.

x	13	14	15	16	17	18	19	20	21	22	23	24
y	-11.6	-17.2	-18.0	-15.0	-8.4	6.2	12.0	20.1	18.4	11.9	3.4	-6.9

A sine function should be used to model the data. We can use a graphing calculator to find the **sinusoidal** regression equation. Enter the x -values into L_1 and the y -values into L_2 . Then with the calculator in radian mode, choose SinReg from the STAT CALC menu:

ENTER: **STAT** **►** **ALPHA** **C** **VARS** **►**
1 **1** **ENTER**



To the nearest thousandth, the sinusoidal regression equation is:

$$y = 19.251 \sin (0.551x + 2.871) - 0.029$$

Press **ZOOM** **9** to graph the scatter plot and the regression equation.

Note: The sinusoidal regression model on the graphing calculator assumes that the x -values are equally spaced and in increasing order. For arbitrary data, you need to give the calculator an estimate of the period. See your calculator manual for details.

The average high temperature of a city is recorded for 14 months. The table below shows this data.

Month	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Temp (°F)	40	48	61	71	81	85	83	77	67	54	41	39	42	49

- Create a scatter plot for the data.
- Find the sinusoidal regression equation for the data with the coefficient and base rounded to three decimal places.
- Predict the average temperature of the city at 15 months. Round to the nearest degree.
- Predict the average temperature of the city at 16 months. Round to the nearest degree.

15-10 INTERPOLATION AND EXTRAPOLATION

Data are usually found for specific values of one of the variables. Often we wish to approximate values not included in the data.

Interpolation

The process of finding a function value between given values is called **interpolation**.

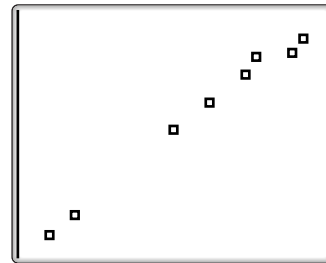
EXAMPLE 1

Each time Jen fills the tank of her car, she records the number of gallons of gas needed to fill the tank and the number of miles driven since the last time that she filled the tank. Her record is shown in the table.

Gallons of Gas	7.5	8.8	5.3	9.0	8.1	4.7	6.9	8.3
Miles	240	280	170	290	260	150	220	270

- If Jen needs 8.0 gallons the next time she fills the tank, to the nearest mile, how many miles will she have driven?
- If Jen has driven 200 miles, to the nearest tenth, how many gallons of gasoline can she expect to need?

Solution Graph the scatter plot of the data. There appears to be positive linear correlation.



With the data in L_1 and L_2 , use the calculator to find the regression equation:

ENTER: **STAT** **►** **4** **ENTER**

When rounded to the nearest thousandth, the linear equation returned is

```
LINREG
Y=AX+B
A=32.36537919
B=-2.076402594
R^2=.9988025622
R=.9994011018
```

$$y = 32.365x - 2.076$$

a. Substitute 8.0 for x in the equation given by the calculator.

$$\begin{aligned} y &= 32.365x - 2.076 \\ &= 32.365(8.0) - 2.076 \\ &= 256.844 \end{aligned}$$

Jen will have driven approximately 257 miles. **Answer**

b. Substitute 200 for y in the equation given by the calculator.

$$\begin{aligned} y &= 32.365x - 2.076 \\ 200 &= 32.365x - 2.076 \\ 202.076 &= 32.365x \\ 6.244 &\approx x \end{aligned}$$

Jen will need approximately 6.2 gallons of gasoline. **Answer**

Extrapolation

Often we want to use data collected about past events to predict the future. The process of using pairs of values within a given range to approximate values outside of the given range of values is called **extrapolation**.

EXAMPLE 2

The following table shows the number of high school graduates in the U.S. in the thousands from 1992 to 2004.

Year	1992	1993	1994	1995	1996	1997	1998
No. of Graduates	2,478	2,481	2,464	2,519	2,518	2,612	2,704
Year	1999	2000	2001	2002	2003	2004	
No. of Graduates	2,759	2,833	2,848	2,906	3,016	3,081	

- Write a linear regression equation for this data.
- If the number of high school graduates continued to grow at this rate, how many graduates would there have been in 2006?
- If the number of high school graduates continues to grow at this rate, when is the number of high school graduates expected to exceed 3.5 million?

Solution a. Enter the year using the number of years since 1990, that is, the difference between the year and 1990, in L_1 . Enter the corresponding number of high school graduates in L_2 . The regression equation is:

$$y = 53.984x + 2,277.286 \quad \text{Answer}$$

- b. Use the equation $y = 53.984x + 2,277.286$ and let $x = 16$:

$$y = 53.984(16) + 2,277.286 \approx 3,141$$

If the increase continued at the same rate, the expected number of graduates in 2006 would have been approximately 3,141,000.

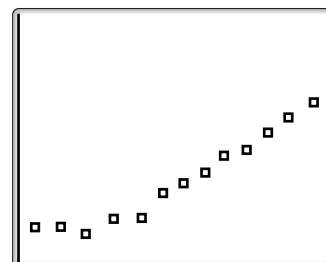
- c. Use the equation $y = 53.984x + 2,277.286$ and let $y = 3,500$:

$$3,500 = 53.984x + 2,277.286$$

$$1,222.714 = 53.984x$$

$$22.650 \approx x$$

If the rate of increase continues, the number of high school graduates can be expected to exceed 3.5 million in the 23rd year after 1990 or in the year 2013.

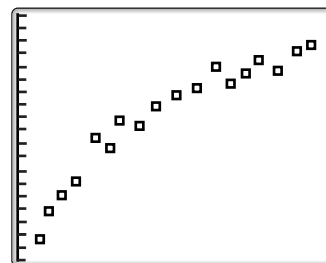


Unlike interpolation, extrapolation is not usually accurate. Extrapolation is valid provided we are sure that the regression model continues to hold outside of the given range of values. Unfortunately, this is not usually the case. For instance, consider the data given in Exercise 15 of Section 15-9.

The growth chart below shows the average height in inches of a group of 100 children from 2 months to 36 months.

Month	2	4	6	8	10	12	14	16	18
Height in Inches	22.7	26.1	27.5	28.9	32.1	31.7	33.1	32.7	34.0
Month	20	22	24	26	28	30	32	34	36
Height in inches	34.4	34.6	36.0	34.6	35.2	36.6	35.6	37.2	37.6

The data appears logarithmic. When the coefficient and the exponent are rounded to three decimal places, an equation that best fits the data is $y = 19.165 + 5.026 \ln x$. If we use this equation to find the height of child who is 16 years old (192 months), the result is approximately 45.6 inches or less than 4 feet. The average 16-year-old is taller than this. The chart is intended to give average growth for very young children and extrapolation beyond the given range of ages leads to errors.



Exercises

Writing About Mathematics

1. Explain the difference between interpolation and extrapolation.
2. What are the possible sources of error when using extrapolation based on the line of best fit?

Developing Skills

In 3–5: **a.** Determine the appropriate linear regression model to use based on the scatter plot of the given data. **b.** Find an approximate value for y for the given value of x . **c.** Find an approximate value for x for the given value of y .

3. **b.** $x = 5.7$ **c.** $y = 1.25$

x	1	2	3	4	5	6	7	8	9	10
y	1.05	1.10	1.16	1.22	1.28	1.34	1.41	1.48	1.55	1.62

4. **b.** $x = 12$ **c.** $y = 140$

x	1	2	3	4	5	6	7	8	9	10
y	3.1	3.6	4.0	4.5	5.1	5.6	6.0	6.5	6.9	7.5

5. **b.** $x = 0.5$ **c.** $y = 0.5$

x	1	2	3	4	5	6	7	8	9	10
y	1	2.3	3.7	5.3	6.9	8.6	10.3	12.1	14.0	15.9

In 6–9: **a.** Determine the appropriate non-linear regression model to use based on the scatter plot of the given data. **b.** Find an approximate value for y for the given value of x . **c.** Find an approximate value for x for the given value of y .

6. **b.** $x = 1.4$ **c.** $y = 1.50$

x	1	2	3	4	5	6	7	8	9	10
y	0.80	1.09	1.25	1.37	1.46	1.54	1.60	1.66	1.71	1.75

7. **a.** $x = 12$ **b.** $y = 80$

x	23	26	13	14	20	17	29	18	18	17
y	11.6	33.3	52.5	43.5	4.0	18.7	84.0	8.0	12.4	11.5

8. **a.** $x = 10.5$ **b.** $y = 100.0$

x	-2.0	-1.0	-0.5	0.1	0.5	0.8	1.1	1.5	1.8	2.1
y	1.0	2.3	3.3	5.3	7.7	9.3	12.0	16.5	21.6	28.0

9. **a.** $x = 12$ **b.** $y = 215$

x	0.5	1.7	2.7	3.9	4.9	5.7	7.0	8.2	9.2	10.0
y	0.1	2.1	7.8	23.1	43.4	65.1	114.9	183.8	248.3	311.3

Applying Skills

In 10–12, determine the appropriate linear regression model to use based on the scatter plot of the given data.

10. The following table represents the percentage of the Gross Domestic Product (GDP) that a country spent on education.

Year	1960	1965	1970	1975	1980	1985	1990	1995	2000	2005
Percent	2.71	3.17	5.20	5.61	7.20	8.14	8.79	10.21	10.72	11.77

- Estimate the percentage of the GDP spent on education in 1998.
 - Assuming this model continues to hold into the future, predict the percentage of the GDP that will be spent on education in 2015.
11. The following chart gives the average time in seconds that a group of 10 F1 racing cars went from zero to the given miles per hour.

Speed	75	100	125	150	175	200	225	250	275	300
Time	1.2	2.2	2.9	4.0	5.6	6.8	7.3	8.7	9.3	9.9

- What was the average time it took the 10 racing cars to reach 180 miles per hour?
 - Estimate the average time it will take the 10 racing cars to reach 325 miles per hour.
12. The relationship between degrees Celsius and degrees Fahrenheit is shown in the table at intervals of 10° Fahrenheit.

Celsius	0	10	20	30	40	50	60	70	80	90	100
Fahrenheit	32	50	68	86	104	122	140	158	176	194	212

- Find the Fahrenheit temperature when the Celsius temperature is 25° .
 - Find the Celsius temperature when the Fahrenheit temperature is -4° .
13. The following table gives the number of compact cars produced in a country over the course of several years.

Year	1981	1984	1987	1990	1993	1996	1999	2002	2005	2008
No. of Cars	100	168	471	603	124	1,780	1,768	4,195	6,680	10,910

- Estimate the number of cars produced by the country in 2000 using an exponential model.
- Estimate the number of cars produced by the country in 1978 using the model from part a.

14. In an office building the thermostats have six settings. The table below shows the average temperature in degrees Fahrenheit for a month that each setting produced.

Setting	1	2	3	4	5	6
Temperature (°F)	61	64	66	67	69	70

Using a power model and assuming that it is possible to choose a setting between the given settings:

- what temperature would result from a setting halfway between 2 and 3?
- where should the setting be placed to produce a temperature of 68 degrees?

In 15 and 16, determine the appropriate non-linear regression model to use based on the scatter plot of the given data.

15. A mail order company has shipping boxes that have square bases and varying height from 1 to 5 feet. The relationship between the height of the box and the volume in cubic feet is shown in the table.

Height	1	1.5	2	2.5	3	3.5	4	4.5	5
Volume	2	6.75	16	31.25	54	85.75	128	182.25	250

- If the company introduces a box with a height of 1.25 feet, what would be the volume to the nearest hundredth cubic foot?
 - If the company needs a box with a volume of at least 100 cubic feet, what would be the smallest height to the nearest tenth of a foot?
 - If the company needs a box with a volume of 800 square feet, what would be the height to the nearest foot?
16. Steve kept a record of the height of a tree that he planted. The heights are shown in the table.

Age of Tree in Years	1	3	5	7	9	11	13
Height in Inches	7	12	15	16.5	17.8	19	20

- Write an equation that best fits the data.
- What was the height of the tree after 2 years?
- If the height of the tree continues in this same pattern, how tall will the tree be after 20 years?

CHAPTER SUMMARY

Univariate Statistics

Statistics is the science that deals with the collection, organization, summarization, and interpretation of related information called **data**. **Univariate statistics** consists of one number for each data value. Data can be collected by means of censuses, surveys, controlled experiments, and observational studies. The source of information in a statistical survey may be the population, all cases to which the study applies, or a sample, a representative subset of the entire group.

The *mean*, the *median*, and the *mode* are the most common measures of **central tendency**. The **mean** is the sum of all of the data values divided by the number of data values.

$$\text{Mean} = \frac{\sum_{i=1}^n f_i x_i}{\sum_{i=1}^n f_i}$$

When finding the mean of data grouped in intervals larger than 1, the median value of each interval is used to represent each entry in the interval.

The **median** is the middle number when the data are arranged in numerical order. In a set of $2n$ data values arranged in numerical order, the median is the average of values that are the n and $(n + 1)$ entries. In a set of $2n + 1$ data values arranged in numerical order, the median is the $n + 1$ entry.

The **mode** is the data value that occurs most frequently.

The **quartiles** separate the data into four equal parts. The **second quartile** is the median. When the data values are arranged in numerical order starting with the smallest value, the **first** or **lower quartile** is the middle value of the values that precede the median, and the **third** or **upper quartile** is the middle value of the values that follow the median.

An **outlier** is a data value that is more than 1.5 times the **interquartile range** above the third quartile or less than 1.5 times the interquartile range below the first quartile.

Measures of Dispersion

- Range = the difference between the smallest and largest data values
- Interquartile range = $Q_3 - Q_1$

$$\text{Variance based on a population} = \frac{\sum_{i=0}^8 f_i (x_i - \bar{x})^2}{\sum_{i=0}^8 f_i}$$

$$\text{Standard deviation based on a population} = \sigma = \sqrt{\frac{\sum_{i=1}^n f_i (x_i - \bar{x})^2}{\sum_{i=1}^n f_i}}$$

$$\bullet \text{ Standard deviation based on a sample } = s = \sqrt{\frac{\sum_{i=1}^n f_i (x_i - \bar{x})^2}{\left(\sum_{i=1}^n f_i\right) - 1}}$$

The Normal Distribution

A **normal curve** represents large sets of data that occur naturally, such as heights, weights, or test grades. The normal curve is bell shaped, symmetric with respect to a vertical line through its highest point, which is at the mean.

A **normal distribution** is a set of data that can be represented by a normal curve. For a normal distribution, the following relationships exist.

1. The mean and the median of the data values lie on the line of symmetry of the curve.
2. Approximately 68% of the data values lie within one standard deviation from the mean.
3. Approximately 95% of the data values lie within two standard deviations from the mean.
4. Approximately 99.7% of the data values lie within three standard deviations from the mean.

The **z-score** for a data value is the deviation from the mean divided by the standard deviation. Let x be a data value of a normal distribution.

$$\text{z-score} = \frac{x - \bar{x}}{\text{standard deviation}} = \frac{x - \bar{x}}{\sigma}$$

The z-score of x , a value from a normal distribution, is positive if x is above the mean and negative if x is below the mean. The z-score tells us how many standard deviations x is above or below the mean.

Bivariate Data

A **scatter plot** can help us to observe the relationship between elements of the pairs of **bivariate data**. The **correlation** between two variables is positive when the values of the second element of the pairs increase when the values of the first elements of the pairs increase. The correlation is negative when the values of the second element of the pairs decrease when the values of the first elements of the pairs increase.

The **correlation coefficient** is a number that indicates the strength and direction of the linear correlation between variables in a set of bivariate data. The correlation coefficient, r , is a number between -1 and 1 . When the absolute value of the correlation coefficient is close to 1 , the data has a strong linear correlation. When the absolute value of the correlation coefficient is close to 0 , there is weak or no linear correlation.

Not all bivariate data can be represented by a linear function. Some data can be better approximated by a curve. By graphing data in a scatter plot, expo-

nential, logarithmic, or power regressions may be identified. A calculator can also be used to find specific linear, exponential, logarithmic, or power functions that best represent the data.

The process of finding a function value between given values is called **interpolation**. The process of using pairs of values within a given range to approximate values outside of the given range of values is called **extrapolation**.

A function value can be found by substituting in the equation of the line of best fit. Extrapolation can sometimes lead to inaccurate results when the extreme values of the data do not fit the pattern of the rest of the data.

VOCABULARY

- 15-1** Statistics • Data • Univariate statistics • Census • Survey • Controlled experiment • Observational study • Population • Sample • Stem-and-leaf diagram • Stem • Leaf • Frequency distribution table • Histogram
- 15-2** Measure of central tendency • Mean • Arithmetic mean • Median • Mode • Quartiles • First quartile • Lower quartile • Second quartile • Third quartile • Upper quartile • Five statistical summary • Box-and-whisker plot
- 15-3** Cumulative frequency • Percentile
- 15-4** Measure of dispersion • Range • Interquartile range • Outlier
- 15-5** Variance • Standard deviation • Standard deviation for a population (σ) • Standard deviation for a sample (s)
- 15-6** Normal curve • Normal distribution • z-score
- 15-7** Bivariate statistics • Scatter plot • Correlation • Regression line • Line of best fit
- 15-8** Correlation coefficient (r)
- 15-9** Sinusoidal
- 15-10** Interpolation • Extrapolation

REVIEW EXERCISES

In 1–3, determine if the data to be collected is univariate or bivariate.

1. The ages of the 20 members of a book club
2. The median heights of boys for each year from age 12 to 18
3. The weight and weight-loss goals of the members of an exercise program

4. Name and describe four common ways of obtaining data for a statistical study.
5. In order to determine the average grade for all students who took a test given to all 9th-grade students in the state, a statistics student at a local college gathered the test grades for five randomly chosen students in the high school of the town where he lived.
- Do the data represent the population or a sample? Explain your answer.
 - Can the person collecting the data expect that the data collected will reflect the grades of all students who took the test? Explain why or why not.
6. Sue's grades are 88, 87, 85, 82, 80, 80, 78, and 60.
- What is the range of her grades?
 - What is the mean grade?
 - What is the median grade?
 - What are the upper and lower quartiles?
 - What is the interquartile range?
 - Is the grade of 60 an outlier? Justify your answer.
 - Draw a box-and-whisker plot for Sue's grades.
7. The hours, x_i , that Peg worked for each of the last 15 weeks are shown in the table at the right. Show your work. In **a** through **h**, you are not allowed to use the **STAT** menu of the calculator.
- Find the mean.
 - Find the median.
 - Find the mode.
 - What is the range?
 - What are the first and third quartiles?
 - What is the interquartile range?
 - Find the variance.
 - Find the standard deviation.
 - Use the **STAT** menu on a calculator and compare the values given with those found in **a**, **b**, **e**, and **h**.

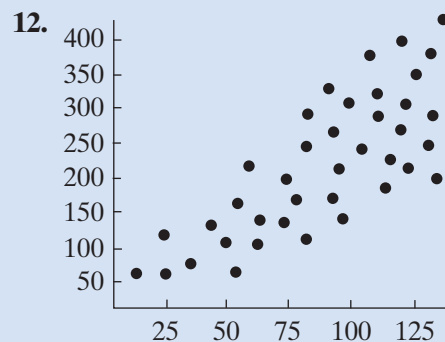
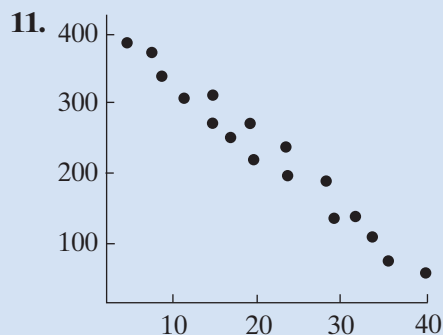
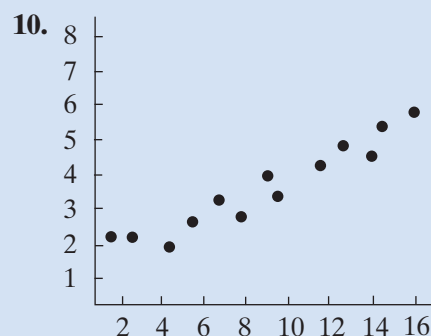
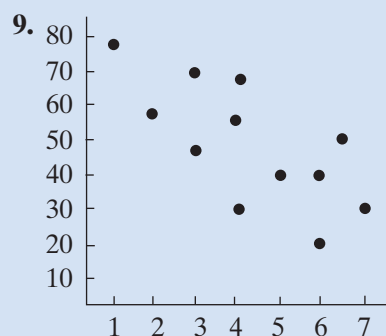
Hours x_i	Frequency f_i
42	1
40	2
39	2
38	5
37	3
36	2

8. Each time Kurt swims, he times a few random laps. His times, in seconds, are shown in the following table.

Time	79	80	81	82	83	84	85	86	87	88	89	90
Frequency	1	3	9	17	30	40	41	29	18	8	3	1

- What percent of the data lie within 1 standard deviation from the mean?
- What percent of the data lie within 2 standard deviations from the mean?
- Does the data appear to represent a normal distribution? Justify your answer.

In 9–12, for each of the given scatter plots: **a.** Describe the correlation as strong positive linear correlation, moderate positive linear correlation, no linear correlation, moderate negative linear correlation, or strong negative linear correlation. **b.** Determine whether the correlation coefficient would be positive or negative.



- 13.** A school investment club chose nine stocks at the beginning of the year that it believed would represent a good investment portfolio. The chart shows the price of the stock at the beginning of the year and the gain (a positive number) or loss (a negative number) at the end of the year.

Stock Price	31.94	18.12	20.76	29.65	16.18	10.14	45.85	56.30	40.15
Gain or Loss	7.19	4.20	-3.57	-2.61	-3.32	-5.75	14.76	9.23	-3.93

- a.** Draw a scatter plot for the data.
- b.** Does there appear to be a linear correlation between the stock price and the gain or loss? If so, is the correlation strong or moderate?

In 14 and 15, find: **a.** the equation of the linear regression model that appears to be appropriate for the data. **b.** the value of the correlation coefficient.

- 14.** The table lists the six states that had the largest percent of increase in population from 2006 to 2007. Population is given in millions, that is, 2.49 represents 2,490,000.

State	Nevada	Arizona	Utah	Idaho	Georgia	N. Carolina
Population 2006	2.49	6.17	2.58	1.46	9.34	8.87
Population 2007	2.57	6.34	2.65	1.50	9.54	9.06

- 15.** The salaries that Aaron earned for four years are shown in the table. The salary is given to the nearest thousand dollars.

Year	1	2	3	4
Salary	52	54	61	63

In 16 and 17, determine the equation of the non-linear regression model that appears to be appropriate for the data.

- 16.** Mrs. Brudek bakes and sells cookies. The table shows the number of dozens of cookies that she baked each week for the first seven weeks of this year.

Week	1	2	3	4	5	6	7
Dozens of Cookies	120	139	162	191	222	257	300

17. In a local park, an attempt is being made to control the deer population. The table shows the estimated number of deer in the park for the six years that the program has been in place.

Year	1	2	3	4	5	6
Number of Deer	700	525	425	350	300	250

In 18–21, use your answers to Exercises 14–17 to estimate the required values.

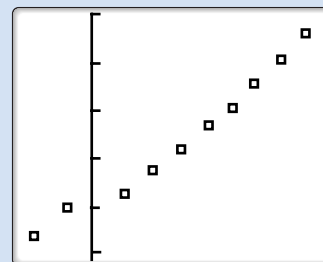
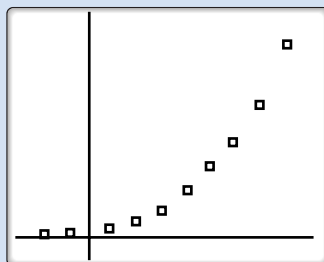
18. Use the information in Exercise 14 to find the approximate population of Texas in 2007 if the population in 2006 was 23.9 million.
19. Use the information in Exercise 15 to predict Aaron's salary in the 5th year.
20. If the number of dozens of cookies that Mrs. Brudek bakes continues to increase according to the pattern shown in Exercise 16, how many dozen cookies will she bake in week 8?
21. a. Use the information in Exercise 17 to find the estimated number of deer in the 8th year of the program if the number of deer continues to decrease according to the given pattern.
- b. Using the information shown in Exercise 17, in what year will the number of deer be approximately 200?

Exploration

The graphing calculator fits a non-linear function to a set of data by transforming the data in such a way that the resulting data fall into a linear pattern. The calculator then fits a *linear* equation to the transformed data and applies an inverse function to find the equation of the non-linear function. In this Exploration, we will be using this procedure.

Look at the following set of data.

x	-2	-1	1	2	3	4	5	6	7	8
y	13	20	29	45	65	114	188	258	378	583
$\ln y$	1.125	1.294	1.504	1.699	1.830	2.058	2.273	2.412	2.578	2.766



The scatter plot on the left reveals that the data is exponential. The scatter plot on the right is a graph of the ordered pairs (x, y') where $y' = \ln y$. Notice

that this scatter plot indicates a clear linear pattern. The data is said to have been **linearized**. If we now find the *linear* regression equation of the ordered pairs (x, y') , the equation is

$$y' = 0.387x + 3.197$$

However, this equation is really

$$\ln y = 0.387x + 3.197$$

To find the exponential regression model, raise both sides to the power e .

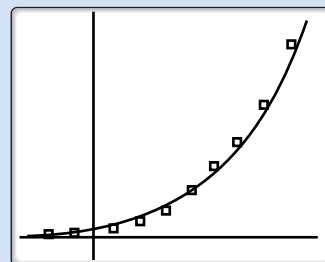
$$\ln y = 0.387x + 3.197$$

$$e^{\ln y} = e^{0.387x + 3.197}$$

$$y = e^{0.387x} \cdot e^{3.197}$$

$$y = (e^{0.387})^x (e^{3.197})$$

$$y = 24.459(1.473)^x$$



The scatter plot on the right shows the original data and the graph of the exponential regression equation.

For other types of non-linear data, we can follow a similar procedure. First we linearize the data using the appropriate substitution(s). Then we find the *linear* regression equation on the transformed data. Lastly, we undo the transformation to obtain the non-linear regression equation.

The substitutions to linearize power and logarithmic data are given below:

Power: $(x, y) \rightarrow (\ln x, \ln y)$ to obtain $\ln y = a \ln x + b$

Logarithmic: $(x, y) \rightarrow (\ln x, y)$ to obtain $y = a \ln x + b$

1. Find a power model using the techniques of this Exploration for the data given in Example 1 of Section 15-9.
2. Find a logarithmic model using the techniques of this Exploration for the data given in Example 2 of Section 15-9.

CUMULATIVE REVIEW

CHAPTERS 1-15

Part I

Answer all questions in this part. Each correct answer will receive 2 credits. No partial credit will be allowed.

1. Solve for x : $2 + \sqrt{x + 3} = 6$

(1) 1

(2) 13

(3) 31

(4) 33

2. An angle measure of 240° is equivalent to
 (1) π radians (3) $\frac{4}{3}\pi$ radians
 (2) $\frac{3}{4}\pi$ radians (4) $\frac{3}{2}\pi$ radians
3. If $f(x) = 3x + 2$ and $g(x) = x^2 - 2$, then $f(g(3))$ is equal to
 (1) 18 (2) 23 (3) 77 (4) 119
4. In the set of real numbers, what is the largest possible domain of the function $f(x) = \frac{2x^2 - 2}{x - 1}$?
 (1) $\{x : x \neq 1\}$ (3) $\{x : x \neq 2\}$
 (2) $\{x : x \neq 1 \text{ and } x \neq -1\}$ (4) $\{x : x \text{ is any real number}\}$
5. The expression $\frac{\sqrt{-36}}{-\sqrt{36}}$ is equal to
 (1) 1 (2) -1 (3) i (4) $-i$
6. If $\sin \theta < 0$ and $\tan \theta = -\frac{4}{5}$, in what quadrant does the terminal side of θ lie when θ is in standard position?
 (1) I (2) II (3) III (4) IV
7. For $a \neq 0, 1, -1$, the expression $\frac{\frac{a-1}{a}}{\frac{a^2-1}{a^2}}$ is equivalent to
 (1) $\frac{a}{a+1}$ (2) $\frac{a+1}{a}$ (3) $\frac{a}{a-1}$ (4) $\frac{a-1}{a}$
8. The roots of the equation $x^2 + 7x - 8 = 0$ are
 (1) real, rational, and equal (3) real, irrational, and unequal
 (2) real, rational, and unequal (4) imaginary
9. The sum of the infinite geometric series when the first term is 8 and the common ratio is 0.2 is
 (1) 0.1 (2) 1 (3) 10 (4) 40
10. The product $(-2 + 6i)(3 + 4i)$ is equal to
 (1) $-6 + 24i$ (3) $18 + 10i$
 (2) $-6 - 24i$ (4) $-30 + 10i$

Part II

Answer all questions in this part. Each correct answer will receive 2 credits. Clearly indicate the necessary steps, including appropriate formula substitutions, diagrams, graphs, charts, etc. For all questions in this part, a correct numerical answer with no work shown will receive only 1 credit.

11. What are the roots of $x^2 - 6x + 13 = 0$?
12. Sketch the graph of $y = 2 \cos \frac{1}{2}x$ in the interval $-2\pi \leq x \leq 2\pi$.

Part III

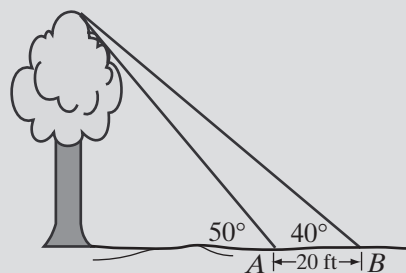
Answer all questions in this part. Each correct answer will receive 4 credits. Clearly indicate the necessary steps, including appropriate formula substitutions, diagrams, graphs, charts, etc. For all questions in this part, a correct numerical answer with no work shown will receive only 1 credit.

13. If $\log 2 = a$ and $\log 3 = b$, express $\log \frac{\sqrt[3]{6}}{9}$ in terms of a and b .
14. Solve for x : $27^{x+1} = 81^x$

Part IV

Answer all questions in this part. Each correct answer will receive 6 credits. Clearly indicate the necessary steps, including appropriate formula substitutions, diagrams, graphs, charts, etc. For all questions in this part, a correct numerical answer with no work shown will receive only 1 credit.

15. A tree service wants to estimate the height of a tree. At point A , the angle of elevation of the top of the tree is 50° . From point B , 20 feet farther from the foot of the tree than point A , the angle of elevation of the top of the tree is 40° . If points A and B lie on a line perpendicular to the tree trunk at its base, what is the height of the tree to the nearest foot?



16. Solve the following system of equations graphically:

$$y = -x^2 + 2x$$

$$y = 2^x - 1$$