B.1   Representing Data

- **Understand terminology associated with statistics.**
- **Use line plots to order and analyze data.**
- **Use stem-and-leaf plots to organize and compare data.**
- **Use histograms to represent frequency distributions.**

An Overview of Statistics

Statistics is the branch of mathematics that studies techniques for collecting, organizing, and interpreting information. **Data** consist of information that comes from observations, responses, counts, or measurements. Sometimes, to make the data easy to understand, they are presented graphically, as illustrated below.
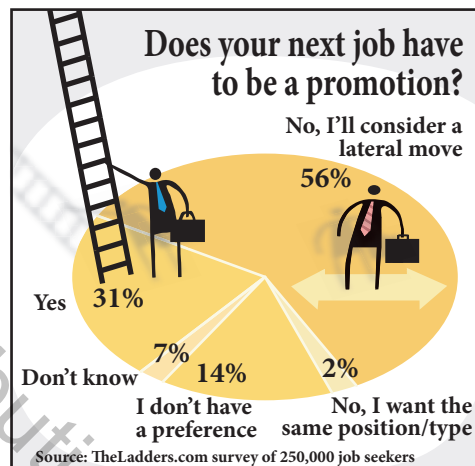
Line plots, stem-and-leaf plots, and histograms are quick methods for determining which elements in a set of data occur with the greatest frequency. For instance, in Exercise 18 on page 880, you will use a line plot to analyze cattle weights.

Does your next job have to be a promotion?

No, I'll consider a lateral move **56%**

Yes **31%**

Don't know **7%**   **14%** I don't have a preference

2% No, I want the same position/type

Source: TheLadders.com survey of 250,000 job seekers

> **REMARK** *Data* is the plural of *datum*.

Two types of data sets that you will use in your study of statistics include *populations* and *samples*. **Populations** are collections of *all* of the outcomes, measurements, counts, or responses that are of interest. **Samples** are subsets, or parts, of a population. A **Venn diagram** is a schematic representation used to depict sets and show how they are related. For instance, suppose that in a recent survey of 1000 U.S. adults, 590 think that purchasing a home is the best investment that a family can make. *(Source: Rasmussen Reports)* The following Venn diagram illustrates the results of this survey.

Responses of all U.S. adults (population)

Responses of adults in survey (sample)
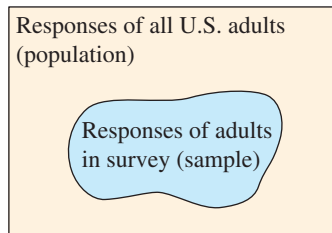
In this example, the population is the collection of the responses of all U.S. adults, and the sample is the collection of the responses of those 1000 U.S. adults surveyed. The sample consists of 590 responses that purchasing a home is the best investment, and 410 responses that purchasing a home is not the best investment (or that some other investment is better). The sample is a part, or subset, of the responses of all U.S. adults.

Two major branches in the study of statistics include *descriptive statistics* and *inferential statistics*. **Descriptive statistics** involves organizing, summarizing, and displaying data. **Inferential statistics** involves drawing conclusions about a population using a sample. The following example illustrates the difference between descriptive and inferential statistics.

EXAMPLE 1 **Descriptive and Inferential Statistics**

A sample of 48-year-old men was studied for 18 years. Approximately 70% of the unmarried men and 90% of the married men were alive at age 65. *(Source: The Journal of Family Issues)*

a. Which part of the study represents descriptive statistics?

b. Use inferential statistics to draw a conclusion from the study.

Solution

a. Descriptive statistics represented in the study include the statement "Approximately 70% of the unmarried men and 90% of the married men were alive at age 65."

b. One possible conclusion you could draw from the study using inferential statistics is that married men tend to live longer than unmarried men.

✓ *Checkpoint* 🔊))) *Audio-video solution in English & Spanish at LarsonPrecalculus.com.*

In a sample of Wall Street analysts, 44% incorrectly forecasted earnings in the high-tech industry in a recent year. *(Source: Bloomberg News)*

a. Which part of the study represents descriptive statistics?

b. Use inferential statistics to draw a conclusion from the study. ◼

Data sets can be made up of two types of data: *quantitative data* and *qualitative data*. **Quantitative data** is made up of numerical values. **Qualitative data** is made up of attributes, labels, or nonnumerical values. For instance, qualitative data can include such elements as eye color, "yes" or "no" responses, or answers given in an opinion poll.

In statistical studies, researchers more commonly use samples rather than performing a **census,** which is a survey of an entire population. Different types of sampling methods are shown below.

Sampling Methods

In a **random sample,** every member of the population has an equal chance of being selected.

In a **stratified random sample,** researchers divide the population into distinct groups and select members at random from each group.

In a **systematic sample,** researchers use a rule to select members of the population. For instance, researchers may select every 4th, 10th, or 1000th member.

In a **convenience sample,** researchers only select members of the population who are easily accessible.

In a **self-selected sample,** members of the population select themselves by volunteering.

> **REMARK** A convenience sample is not recommended because it often leads to biased results.

In the remainder of this section, you will study several ways to organize data. The first is a *line plot*.

Line Plots

A **line plot** uses a portion of a real number line to order numbers. Line plots are especially useful for ordering small sets of numbers (about 50 or less) by hand.

EXAMPLE 2   **Constructing a Line Plot**

Use a line plot to organize the following test scores. Which score occurs with the greatest frequency?   *(Spreadsheet at LarsonPrecalculus.com)*

DATA   93, 70, 76, 67, 86, 93, 82, 78, 83, 86, 64, 78, 76, 66, 83
83, 96, 74, 69, 76, 64, 74, 79, 76, 88, 76, 81, 82, 74, 70

Solution   Begin by scanning the data to find the least and greatest numbers. For this data set, the least number is 64 and the greatest is 96. Next, draw a portion of a real number line that includes the interval $[64, 96]$. To create the line plot, start with the first number, 93, and enter a ● above 93 on the number line. Continue recording ●'s for the numbers in the list until you obtain the line plot shown below. From the line plot, you can see that 76 occurs with the greatest frequency.


Test scores

✓ **Checkpoint** 🔊)) *Audio-video solution in English & Spanish at LarsonPrecalculus.com.*

Use a line plot to organize the following test scores. Which score occurs with the greatest frequency?   *(Spreadsheet at LarsonPrecalculus.com)*

DATA   68, 73, 67, 95, 71, 82, 85, 74, 82, 61, 87, 92, 78, 74, 64,
71, 74, 82, 71, 83, 92, 82, 78, 72, 82, 64, 85, 67, 71, 62

EXAMPLE 3   **Analyzing a Line Plot**

The line plot shows the high temperatures (in degrees Fahrenheit) in a city for each day in the month of June.

High temperatures
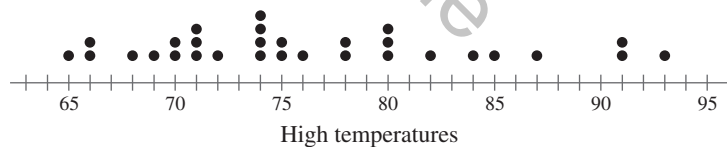
a. What is the range of high temperatures?

b. On how many days was the high temperature in the 80s?

Solution

a. The line plot shows that the maximum high temperature was 93°F and the minimum high temperature was 65°F. So, the range is $93 - 65 = 28$°F.

b. There are 7 ●'s in the interval $[80, 90)$. So, the high temperature was in the 80s on 7 days.

✓ **Checkpoint** 🔊)) *Audio-video solution in English & Spanish at LarsonPrecalculus.com.*

In Example 3, on how many days was the high temperature less than 80°F?

Stem-and-Leaf Plots

Another type of plot that can enable you to organize sets of numbers by hand is a **stem-and-leaf plot.** A set of test scores and the corresponding stem-and-leaf plot are shown below.

Test Scores

93, 70, 76, 58, 86, 93, 82, 78,
83, 86, 64, 78, 76, 66, 83, 83,
96, 74, 69, 76, 64, 74, 79, 76,
88, 76, 81, 82, 74, 70

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

Key: 5|8 = 58

Note that the *leaves* represent the units digits of the numbers and the *stems* represent the tens digits. Stem-and-leaf plots can also help you compare two sets of data. For instance, suppose you want to compare the above test scores with the following test scores.

90, 81, 70, 62, 64, 73, 81, 92, 73, 81,
92, 93, 83, 75, 76, 83, 94, 96, 86, 77,
77, 86, 96, 86, 77, 86, 87, 87, 79, 88

Begin by ordering the second set of scores.

62, 64, 70, 73, 73, 75, 76, 77, 77, 77,
79, 81, 81, 81, 83, 83, 86, 86, 86, 86,
87, 87, 88, 90, 92, 92, 93, 94, 96, 96

Now that the data have been ordered, you can construct a *double* stem-and-leaf plot by letting the leaves to the right of the stems represent the units digits for the first group of test scores and letting the leaves to the left of the stems represent the units digits for the second group of test scores.

Leaves (2nd Group)	Stems	Leaves (1st Group)
	5	8
4 2	6	4 4 6 9
9 7 7 7 6 5 3 3 0	7	0 0 4 4 4 6 6 6 6 8 8 9
8 7 7 6 6 6 6 3 3 1 1 1	8	1 2 2 3 3 3 6 6 8
6 6 4 3 2 2 0	9	3 3 6

Key: 2|6|4 = 64 for 1st group, 62 for 2nd group

By comparing the two sets of leaves, you can see that the second group of test scores is better than the first group.

EXAMPLE 4   **Using a Stem-and-Leaf Plot**

The table at the left shows the percent of the population of each state, and the District of Columbia, that was at least 65 years old in 2010. Use a stem-and-leaf plot to organize the data. *(Source: U.S. Census Bureau)*

Solution   Begin by ordering the numbers. Next, construct the stem-and-leaf plot by letting the leaves represent the digits to the right of the decimal points, as shown at the left. From the stem-and-leaf plot, you can see that Alaska has the lowest percent and Florida has the highest percent.

✓ *Checkpoint* ◀))) *Audio-video solution in English & Spanish at LarsonPrecalculus.com.*

Use the stem-and-leaf plot in Example 4 to determine how many states had populations in which 14% or more people were at least 65 years old in 2010.

Spreadsheet at LarsonPrecalculus.com

DATA

AK	7.7	MT	14.8
AL	13.8	NC	12.9
AR	14.4	ND	14.5
AZ	13.8	NE	13.5
CA	11.4	NH	13.5
CO	10.9	NJ	13.5
CT	14.2	NM	13.2
DC	11.4	NV	12.0
DE	14.4	NY	13.5
FL	17.3	OH	14.1
GA	10.7	OK	13.5
HI	14.3	OR	13.9
IA	14.9	PA	15.4
ID	12.4	RI	14.4
IL	12.5	SC	13.7
IN	13.0	SD	14.3
KS	13.2	TN	13.4
KY	13.3	TX	10.3
LA	12.3	UT	9.0
MA	13.8	VA	12.2
MD	12.3	VT	14.6
ME	15.9	WA	12.3
MI	13.8	WI	13.7
MN	12.9	WV	16.0
MO	14.0	WY	12.4
MS	12.8		

Stems	Leaves
7	7
8	
9	0
10	3 7 9
11	4 4
12	0 2 3 3 3 4 4 5 8 9 9
13	0 2 2 3 4 5 5 5 5 5 7 7 8 8 8 8 9
14	0 1 2 3 3 4 4 4 5 6 8 9
15	4 9
16	0
17	3

Key: 7|7 = 7.7

Histograms and Frequency Distributions

With data such as that given in Example 4, it is useful to group the numbers into intervals and plot the frequency of the data in each interval. For instance, the **frequency distribution** and **histogram** shown below represent the data given in Example 4.

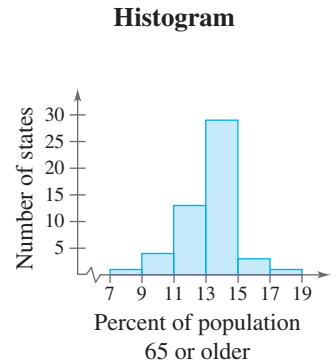| Frequency Distribution		Histogram
Interval	*Tally*	
[7, 9)		
[9, 11)	||||
[11, 13)	$\cancel{||||}$ $\cancel{||||}$ |||
[13, 15)	$\cancel{||||}$ $\cancel{||||}$ $\cancel{||||}$ $\cancel{||||}$ $\cancel{||||}$ ||||
[15, 17)	|||
[17, 19)		

Histograms and frequency distributions have a wide variety of real-life applications. For instance, in Exercise 21, you will use a histogram and a frequency distribution to organize data on employee contributions to a retirement plan.

A histogram has a portion of a real number line as its horizontal axis. A histogram is similar to a bar graph, except that the rectangles (bars) in a bar graph can be either horizontal or vertical and the labels of the bars are not necessarily numbers. Another difference between a bar graph and a histogram is that the bars in a bar graph are usually separated by spaces, whereas the bars in a histogram are not separated by spaces.

EXAMPLE 5   Constructing a Histogram

A company has 48 sales representatives who sold the following numbers of units during the last quarter of 2013. Construct a frequency distribution and histogram for this data set.   *(Spreadsheet at LarsonPrecalculus.com)*

DATA

107	162	184	170	177	102	145	141	105	193	167	149
195	127	193	191	150	153	164	167	171	163	141	129
109	171	150	138	100	164	147	153	171	163	118	142
107	144	100	132	153	107	124	162	192	134	187	177

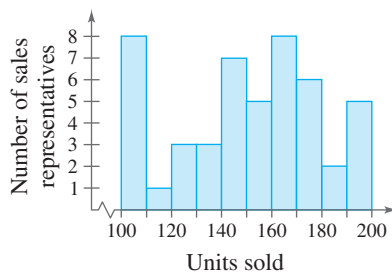**Solution**   To construct a frequency distribution, first decide on the number of intervals. Because the least number is 100 and the greatest is 195, it seems that 10 intervals would be appropriate. The first interval would be 100–109, the second interval would be 110–119, and so on. By tallying the data into the 10 intervals, you obtain the distribution shown at the left. Figure B.1 shows a histogram for the distribution.

✓ *Checkpoint* 🔊))) *Audio-video solution in English & Spanish at LarsonPrecalculus.com.*

Construct a frequency distribution and histogram for the data set in Example 2. ■

Interval	*Tally*
100–109	$\cancel{||||}$ |||
110–119		
120–129	|||
130–139	|||
140–149	$\cancel{||||}$ ||
150–159	$\cancel{||||}$
160–169	$\cancel{||||}$ |||
170–179	$\cancel{||||}$	
180–189	||
190–199	$\cancel{||||}$

Figure B.1

Summarize   (Section B.1)

1. State the definitions of some terminology associated with statistics *(pages B1 and B2)*.

2. Explain how to construct a line plot *(page B3)*. For examples of using line plots to order and analyze data, see Examples 2 and 3.

3. Explain how to use a stem-and-leaf plot to organize and compare data *(page B4)*. For an example of using a stem-and-leaf plot, see Example 4.

4. Explain how to construct a frequency distribution and a histogram *(page B5)*. For an example of using a histogram to represent a frequency distribution, see Example 5.
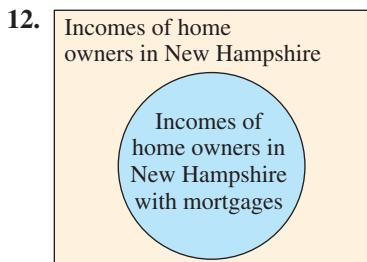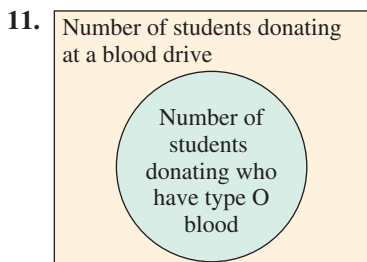
B.1 Exercises

Vocabulary: Fill in the blanks.

1. _____ is the branch of mathematics that studies techniques for collecting, organizing, and interpreting information.

2. _____ are collections of *all* of the outcomes, measurements, counts, or responses that are of interest.

3. _____ are subsets, or parts, of a population.

4. _____ statistics involves organizing, summarizing, and displaying data, whereas _____ statistics involves drawing conclusions about a population using a sample.

5. _____ data is made up of numerical values, whereas _____ data is made up of attributes, labels, or nonnumerical values.

6. In a _____ sample, every member of the population has an equal chance of being selected.

7. In a _____ _____ sample, researchers divide the population into distinct groups and select members at random from each group.

8. In a _____ sample, which is not recommended because it often leads to biased results, researchers only select members of the population who are easily accessible.

9. In a _____ sample, members of the population select themselves by volunteering.

10. A _____ has a portion of a real number line as its horizontal axis, and the bars are not separated by spaces.

Skills and Applications

Venn Diagrams **In Exercises 11 and 12, use the Venn diagram to determine the population and the sample.**

11.
Number of students donating at a blood drive

Number of students donating who have type O blood

12.
Incomes of home owners in New Hampshire

Incomes of home owners in New Hampshire with mortgages

13. **Physical Examinations** A survey conducted among 1017 adults by Opinion Research Corporation International determined that 60% of men and 76% of women had received physical examinations within the past year. *(Source: Men's Health)*

(a) Which part of the study represents descriptive statistics?

(b) Use inferential statistics to draw a conclusion from the study.

14. **Vacation Spending** A survey conducted among 791 U.S. vacationers determined that the vacationers are planning to spend at least $2000 on their next vacations.

(a) Which part of the study represents descriptive statistics?

(b) Use inferential statistics to draw a conclusion from the study.

Quiz and Exam Scores **In Exercises 15 and 16, use the following scores from a math class of 30 students. The scores are for two 25-point quizzes and two 100-point exams.** *(Spreadsheet at LarsonPrecalculus.com)*

DATA

Quiz #1 20, 15, 14, 20, 16, 19, 10, 21, 24, 15, 15, 14, 15, 21, 19, 15, 20, 18, 18, 22, 18, 16, 18, 19, 21, 19, 16, 20, 14, 12

Quiz #2 22, 22, 23, 22, 21, 24, 22, 19, 21, 23, 23, 25, 24, 22, 22, 23, 23, 23, 23, 22, 24, 23, 22, 24, 21, 24, 16, 21, 16, 14

Exam #1 77, 100, 77, 70, 83, 89, 87, 85, 81, 84, 81, 78, 89, 78, 88, 85, 90, 92, 75, 81, 85, 100, 98, 81, 78, 75, 85, 89, 82, 75

Exam #2 76, 78, 73, 59, 70, 81, 71, 66, 66, 73, 68, 67, 63, 67, 77, 84, 87, 71, 78, 78, 90, 80, 77, 70, 80, 64, 74, 68, 68, 68

15. Construct a line plot for each quiz. For each quiz, which score occurred with the greatest frequency?

16. Construct a line plot for each exam. For each exam, which score occurred with the greatest frequency?

17. Gasoline Prices The line plot shows a sample of prices of unleaded regular gasoline from 25 different cities.
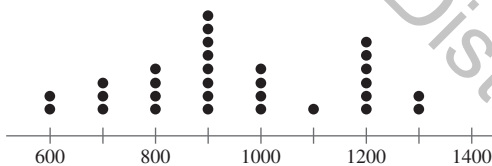

3.699 3.719 3.739 3.759 3.779 3.799 3.819 3.839 3.859 3.879 3.899
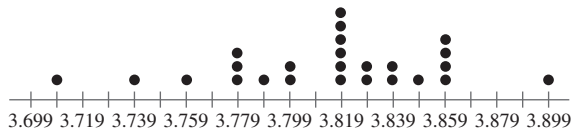
(a) What price occurred with the greatest frequency?

(b) What is the range of prices?

18. Cattle Weights

The line plot shows the weights (to the nearest hundred pounds) of 30 cattle sold by a rancher.

(a) What weight occurred with the greatest frequency?

(b) What is the range of weights?

600   800   1000   1200   1400

Exam Scores In Exercises 19 and 20, use the following scores from a math class of 30 students. The scores are for two 100-point exams. *(Spreadsheet at LarsonPrecalculus.com)*

Exam #1 77, 100, 77, 70, 83, 89, 87, 85, 81, 84, 81, 78, 89, 78, 88, 85, 90, 92, 75, 81, 85, 100, 98, 81, 78, 75, 85, 89, 82, 75

Exam #2 76, 78, 73, 59, 70, 81, 71, 66, 66, 73, 68, 67, 63, 67, 77, 84, 87, 71, 78, 78, 90, 80, 77, 70, 80, 64, 74, 68, 68, 68

19. Construct a stem-and-leaf plot for Exam #1.

20. Construct a double stem-and-leaf plot to compare the scores for Exam #1 and Exam #2. Which set of scores is higher?

21. Retirement Contributions The amounts (in dollars) 35 employees contribute to a retirement plan are shown below. Use a frequency distribution and a histogram to organize the data. *(Spreadsheet at LarsonPrecalculus.com)*

100, 200, 130, 136, 161, 156, 209, 126, 135, 98, 114, 117, 168, 133, 140, 124, 172, 127, 143, 157, 124, 152, 104, 126, 155, 92, 194, 115, 120, 136, 148, 112, 116, 146, 96

22. Agriculture The table shows the total numbers of farms (in thousands) in the 50 states in 2010. Use a frequency distribution and a histogram to organize the data. *(Source: U.S. Dept. of Agriculture)* *(Spreadsheet at LarsonPrecalculus.com)*

AK	1	AL	49	AR	49	AZ	16	CA	82
CO	36	CT	5	DE	2	FL	48	GA	47
HI	8	IA	92	ID	26	IL	76	IN	62
KS	66	KY	86	LA	30	MA	8	MD	13
ME	8	MI	55	MN	81	MO	108	MS	42
MT	29	NC	52	ND	32	NE	47	NH	4
NJ	10	NM	21	NV	3	NY	36	OH	75
OK	87	OR	39	PA	63	RI	1	SC	27
SD	32	TN	78	TX	248	UT	17	VA	47
VT	7	WA	40	WI	78	WV	23	WY	11

23. Meteorology The data show the seasonal snowfall amounts (in inches) for Chicago, Illinois, for the years 1982 through 2011 (the amounts are listed in order by year). How would you organize the data? Explain your reasoning. *(Source: National Weather Service)* *(Spreadsheet at LarsonPrecalculus.com)*

26.6, 49.0, 39.1, 29.0, 26.2, 42.6, 24.5, 33.8, 36.7, 28.4, 46.9, 41.8, 24.1, 23.9, 40.6, 29.6, 50.9, 30.3, 39.2, 31.1, 28.6, 24.8, 39.4, 26.0, 35.6, 60.3, 52.7, 54.2, 57.9, 19.8
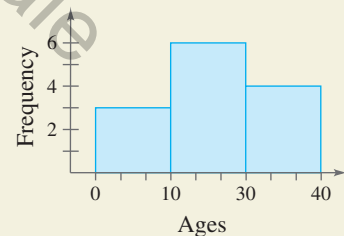
Exploration

24. HOW DO YOU SEE IT?

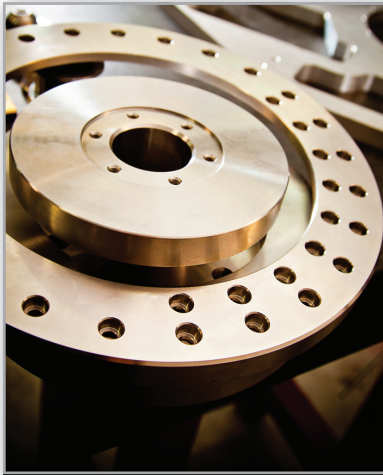Describe and correct the error in creating the histogram below using the frequency distribution at the right.

Interval	Tally
[0, 10)	|||
[10, 20)	|||||	
[20, 30)	
[30, 40)	||||

True or False? In Exercises 25 and 26, determine whether the statement is true or false. Justify your answer.

25. A census surveys an entire population.

26. Using a systematic sample will guarantee the sampling of members of each group in a population.

B.2   Analyzing Data

- Find and interpret the mean, median, and mode of a set of data.
- Determine the measure of central tendency that best represents a set of data.
- Find the standard deviation of a set of data.
- Create and use box-and-whisker plots.
- Interpret normally distributed data.

Measures of central tendency and dispersion provide a convenient way to describe and compare sets of data. For instance, in Exercises 41 and 42 on page B18, you will use box-and-whisker plots to analyze the lifetime of a machine part.

Mean, Median, and Mode

In many real-life situations, it is helpful to describe a data set by a single number that is most representative of the entire collection of numbers. Such a number is called a **measure of central tendency.** The most commonly used measures are as follows.

1. The **mean,** or **average,** of n numbers is the sum of the numbers divided by n.
2. The numerical **median** of n numbers is the middle number when the numbers are written in order. When n is even, the median is the average of the two middle numbers.
3. The **mode** of n numbers is the number that occurs most frequently. When two numbers tie for most frequent occurrence, the collection has two modes and is called **bimodal.**

EXAMPLE 1   Comparing Measures of Central Tendency

An analyst states that the average annual income of a company's 25 employees is $60,849. The actual annual incomes of the 25 employees are shown below. What are the mean, median, and mode of the incomes?   *(Spreadsheet at LarsonPrecalculus.com)*

DATA

$17,305,	$478,320,	$45,678,	$18,980,	$17,408,
$25,676,	$28,906,	$12,500,	$24,540,	$33,450,
$12,500,	$33,855,	$37,450,	$20,432,	$28,956,
$34,983,	$36,540,	$250,921,	$36,853,	$16,430,
$32,654,	$98,213,	$48,980,	$94,024,	$35,671

Solution   The mean of the incomes is

$$\text{Mean} = \frac{17{,}305 + 478{,}320 + 45{,}678 + 18{,}980 + \cdots + 35{,}671}{25} = \$60{,}849.$$

To find the median, order the incomes as follows.

$12,500,	$12,500,	$16,430,	$17,305,	$17,408,
$18,980,	$20,432,	$24,540,	$25,676,	$28,906,
$28,956,	$32,654,	$33,450,	$33,855,	$34,983,
$35,671,	$36,540,	$36,853,	$37,450,	$45,678,
$48,980,	$94,024,	$98,213,	$250,921,	$478,320

From this list, you can see that the median income is $33,450. You can also see that $12,500 is the only income that occurs more than once. So, the mode is $12,500.

✓ *Checkpoint* ◀))) *Audio-video solution in English & Spanish at LarsonPrecalculus.com.*

Find the mean, median, and mode of the following data set.

68, 73, 67, 95, 71, 82, 85, 74, 82, 61

In Example 1, was the analyst telling the truth about the annual incomes? Technically, the analyst was telling the truth because the average is (generally) defined to be the mean. However, of the three measures of central tendency, the median is most representative of this data set.

Choosing a Measure of Central Tendency

Which of the three measures of central tendency is the most representative of a data set? The answer depends on the distribution of the data *and* the way in which you plan to use the data.

For instance, in Example 1, the mean salary of $60,849 does not seem very representative to a potential employee. The two highest salaries inflate the mean. To a city income tax collector who wants to estimate 1% of the mean annual income of the 25 employees, however, the mean is precisely the right measure.

EXAMPLE 2   Choosing a Measure of Central Tendency

Which measure of central tendency is the most representative of the data shown in each frequency distribution?

a.

Number	1	2	3	4	5	6	7	8	9
Frequency	7	20	15	11	8	3	2	0	15

b.

Number	1	2	3	4	5	6	7	8	9
Frequency	9	8	7	6	5	6	7	8	9

c.

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

Solution

a. For these data, the mean is approximately 4.23, the median is 3, and the mode is 2. Of these, the mode is probably the most representative measure.

b. For these data, the mean and median are each 5, and the modes are 1 and 9 (the distribution is bimodal). Of these, the mean or median is the most representative measure.

c. For these data, the mean is approximately 4.59, the median is 5, and the mode is 1. Of these, the mean or median is the most representative measure.

✓ **Checkpoint** 🔊 *Audio-video solution in English & Spanish at LarsonPrecalculus.com.*

Which measure of central tendency is the most representative of the data shown in each frequency distribution?

a.

Number	1	2	3	4	5	6	7	8	9
Frequency	1	1	1	1	3	1	1	1	1

b.

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

c.

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

d.

Number	1	2	3	4	5	6	7	8	9
Frequency	11	8	3	2	0	15	7	20	15

Variance and Standard Deviation

Very different sets of numbers can have the same mean. You will now study two **measures of dispersion,** which give you an idea of how much the numbers in a data set differ from the mean of the set. These two measures are called the *variance* of the set and the *standard deviation* of the set.

Definitions of Variance and Standard Deviation

Consider a set of numbers $\{x_1, x_2, \ldots, x_n\}$ with a mean of $\bar{x}$. The **variance** of the set is

$$v = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n}$$

and the **standard deviation** of the set is $\sigma = \sqrt{v}$ (σ is the lowercase Greek letter *sigma*).

The standard deviation of a data set is a measure of how much a typical number in the set differs from the mean. The greater the standard deviation, the more the numbers in the set *vary* from the mean. For instance, each of the following data sets has a mean of 5.

$$\{5, 5, 5, 5\}, \quad \{4, 4, 6, 6\}, \quad \text{and} \quad \{3, 3, 7, 7\}$$

The standard deviations of the data sets are 0, 1, and 2, respectively.

$$\sigma_1 = \sqrt{\frac{(5 - 5)^2 + (5 - 5)^2 + (5 - 5)^2 + (5 - 5)^2}{4}} = 0$$

$$\sigma_2 = \sqrt{\frac{(4 - 5)^2 + (4 - 5)^2 + (6 - 5)^2 + (6 - 5)^2}{4}} = 1$$

$$\sigma_3 = \sqrt{\frac{(3 - 5)^2 + (3 - 5)^2 + (7 - 5)^2 + (7 - 5)^2}{4}} = 2$$
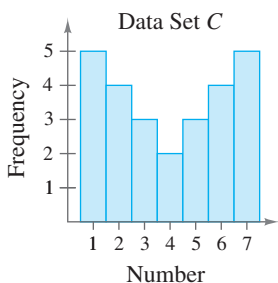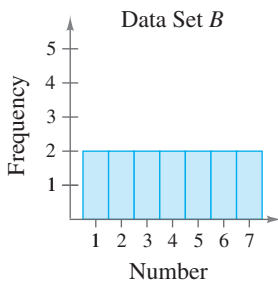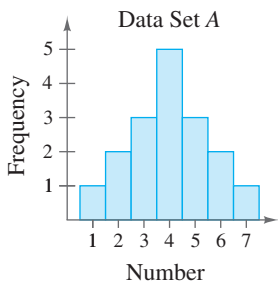
EXAMPLE 3   **Estimations of Standard Deviation**

Consider the three frequency distributions represented by the histograms in Figure B.2. Which data set has the least standard deviation? Which has the greatest?

Solution   Of the three data sets, the numbers in data set *A* are grouped most closely to the center and the numbers in data set *C* are the most dispersed. So, data set *A* has the least standard deviation and data set *C* has the greatest standard deviation.

✓ *Checkpoint* 🔊 *Audio-video solution in English & Spanish at LarsonPrecalculus.com.*

Consider the three frequency distributions represented by the histograms shown below. Which data set has the least standard deviation? Which has the greatest?
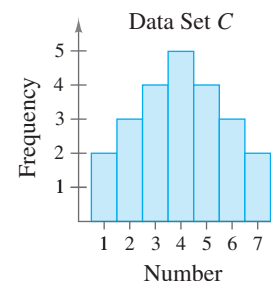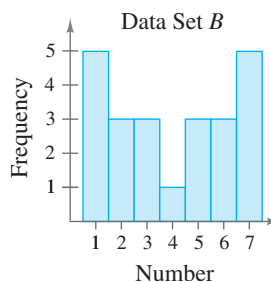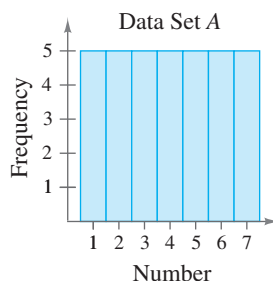

Figure B.2

 EXAMPLE 4 **Finding Standard Deviation**

Find the standard deviation of each data set shown in Example 3.

Solution Because of the symmetry of each histogram, each has a mean of $\bar{x} = 4$. The standard deviation of data set A is

$$\sigma = \sqrt{\frac{(-3)^2 + 2(-2)^2 + 3(-1)^2 + 5(0)^2 + 3(1)^2 + 2(2)^2 + (3)^2}{17}}$$

$$\approx 1.53.$$

The standard deviation of data set B is

$$\sigma = \sqrt{\frac{2(-3)^2 + 2(-2)^2 + 2(-1)^2 + 2(0)^2 + 2(1)^2 + 2(2)^2 + 2(3)^2}{14}}$$

$$= 2.$$

The standard deviation of data set C is

$$\sigma = \sqrt{\frac{5(-3)^2 + 4(-2)^2 + 3(-1)^2 + 2(0)^2 + 3(1)^2 + 4(2)^2 + 5(3)^2}{26}}$$

$$\approx 2.22.$$

These values confirm the results of Example 3. That is, data set A has the least standard deviation and data set C has the greatest.

✓ **Checkpoint** ◀))) *Audio-video solution in English & Spanish at LarsonPrecalculus.com.*

Find the standard deviation of each data set shown in the Checkpoint with Example 3.

■

The following alternative formula provides a more efficient way to compute the standard deviation.

Alternative Formula for Standard Deviation

The standard deviation of $\{x_1, x_2, \ldots, x_n\}$ is

$$\sigma = \sqrt{\frac{x_1^2 + x_2^2 + \cdots + x_n^2}{n} - \bar{x}^2}.$$

Conceptually, the process of proving this formula is straightforward. It consists of showing that the expressions

$$\sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n}}$$

and

$$\sqrt{\frac{x_1^2 + x_2^2 + \cdots + x_n^2}{n} - \bar{x}^2}$$

are equivalent. Try verifying this equivalence for the set $\{x_1, x_2, x_3\}$ with

$$\bar{x} = \frac{x_1 + x_2 + x_3}{3}.$$

For a proof of the more general case, see Proofs in Mathematics on page B32.

EXAMPLE 5 **Using the Alternative Formula**

Use the alternative formula for standard deviation to find the standard deviation of the following set of numbers.

5, 6, 6, 7, 7, 8, 8, 8, 9, 10

Solution Begin by finding the mean of the data set, which is 7.4. So, the standard deviation is

$$\sigma = \sqrt{\frac{5^2 + 2(6^2) + 2(7^2) + 3(8^2) + 9^2 + 10^2}{10} - (7.4)^2}$$

$$= \sqrt{\frac{568}{10} - 54.76}$$

$$\approx 1.43.$$

Check this result using the *one-variable statistics* feature of a graphing utility.

✓ *Checkpoint* 🔊 *Audio-video solution in English & Spanish at LarsonPrecalculus.com.*

Use the alternative formula for standard deviation to find the standard deviation of the following set of numbers.

3, 3, 3, 4, 4, 5, 5, 6, 6, 7

A well-known theorem in statistics, called *Chebychev's Theorem*, states that at least

$$1 - \frac{1}{k^2}$$

of the numbers in a distribution must lie within k standard deviations of the mean. So, at least 75% of the numbers in a data set must lie within two standard deviations of the mean, and at least about 88.9% of the numbers must lie within three standard deviations of the mean. For most distributions, these percentages are low. For instance, in all three distributions shown in Example 3, 100% of the numbers lie within two standard deviations of the mean.

EXAMPLE 6 **Describing a Distribution**

The table at the left shows the number of newspapers published in each state, and the District of Columbia, in 2010. Find the mean and standard deviation of the data. What percent of the data values lie within two standard deviations of the mean? *(Source: Editor & Publisher Co.)*

Solution Begin by entering the numbers into a graphing utility. Then use the *one-variable statistics* feature to obtain $\bar{x} \approx 27.4$ and $\sigma \approx 21.9$. The interval that contains all numbers that lie within two standard deviations of the mean is

$$[27.4 - 2(21.9), 27.4 + 2(21.9)] \quad \text{or} \quad [-16.4, 71.2].$$

From the table, you can see that all but four of the data values (92%) lie in this interval—all but the data values that correspond to the numbers of newspapers published in California, Ohio, Pennsylvania, and Texas.

✓ *Checkpoint* 🔊 *Audio-video solution in English & Spanish at LarsonPrecalculus.com.*

In Example 6, what percent of the data values lie within one standard deviation of the mean?

DATA

Spreadsheet at LarsonPrecalculus.com

AK	7	MT	11
AL	24	NC	46
AR	25	ND	10
AZ	16	NE	15
CA	83	NH	9
CO	29	NJ	17
CT	17	NM	17
DC	2	NV	6
DE	1	NY	60
FL	37	OH	82
GA	34	OK	37
HI	4	OR	18
IA	37	PA	80
ID	11	RI	6
IL	63	SC	16
IN	67	SD	11
KS	35	TN	26
KY	22	TX	81
LA	24	UT	6
MA	32	VA	26
MD	10	VT	9
ME	7	WA	22
MI	48	WI	33
MN	25	WV	20
MO	42	WY	9
MS	22		

Box-and-Whisker Plots

Standard deviation is the measure of dispersion that is associated with the mean. **Quartiles** measure dispersion associated with the median.

Definition of Quartiles

Consider an ordered set of numbers whose median is m. The **lower quartile** is the median of the numbers that occur before m. The **upper quartile** is the median of the numbers that occur after m.

EXAMPLE 7   **Finding Quartiles of a Data Set**

Find the lower and upper quartiles for the following data set.

42, 14, 24, 16, 12, 18, 20, 24, 16, 26, 13, 27

Solution   Begin by ordering the data.

12, 13, 14,	16, 16, 18,	20, 24, 24,	26, 27, 42
1st 25%	2nd 25%	3rd 25%	4th 25%

The median of the entire data set is 19. The median of the six numbers that are less than 19 is 15. So, the lower quartile is 15. The median of the six numbers that are greater than 19 is 25. So, the upper quartile is 25.

✓ *Checkpoint* ◀))) *Audio-video solution in English & Spanish at LarsonPrecalculus.com.*

Find the lower and upper quartiles for the following data set.

39, 47, 81, 43, 23, 23, 27, 86, 15, 3, 74, 55

The **interquartile range** of a data set is the difference of the upper quartile and the lower quartile. In the data set given in Example 7, the interquartile range is $25 - 15 = 10$.

A value that is widely separated from the rest of the data in a data set is called an **outlier.** Typically, a data value is considered to be an outlier when it is greater than the upper quartile by more than 1.5 times the interquartile range or when it is less than the lower quartile by more than 1.5 times the interquartile range. Verify that, in Example 7, the value 42 is an outlier.

Quartiles are represented graphically by a **box-and-whisker plot,** as shown below. In the plot, notice that five numbers are listed: the least number, the lower quartile, the median, the upper quartile, and the greatest number. These numbers are called the **five-number summary** of the data set. Also notice that the numbers are spaced proportionally, as though they were on a real number line.

12	15	19	25	42
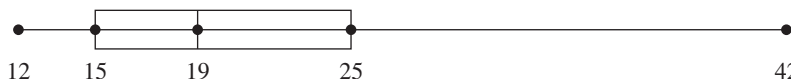
The next example shows how to find quartiles when the number of elements in a data set is not divisible by 4.

▷ **TECHNOLOGY** Some graphing utilities have the capability of creating box-and-whisker plots. If your graphing utility has this capability, then use the data set given in Example 7 to recreate the box-and-whisker plot shown above.
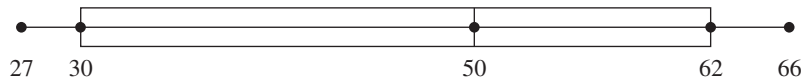
EXAMPLE 8 **Sketching Box-and-Whisker Plots**

Sketch a box-and-whisker plot for each data set.

a. 27, 28, 30, 42, 45, 50, 50, 61, 62, 64, 66

b. 82, 82, 83, 85, 87, 89, 90, 94, 95, 95, 96, 98, 99

**c.** 11, 13, 13, 15, 17, 18, 20, 24, 24, 27

Solution

a. The median is 50. The lower quartile is 30 (the median of the first five numbers). The upper quartile is 62 (the median of the last five numbers). See below.

27   30   50   62   66

b. The median is 90. The lower quartile is 84 (the median of the first six numbers). The upper quartile is 95.5 (the median of the last six numbers). See below.

82   84   90   95.5   99
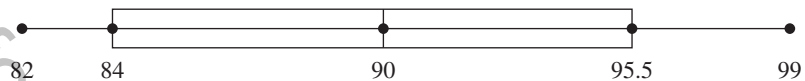
c. The median is 17.5. The lower quartile is 13 (the median of the first five numbers). The upper quartile is 24 (the median of the last five numbers). See below.

11   13   17.5   24   27
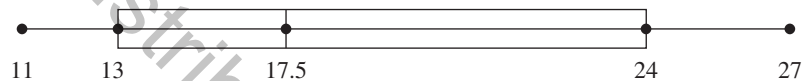
✓ **Checkpoint** 🔊 **Audio-video solution in English & Spanish at LarsonPrecalculus.com.**

Sketch a box-and-whisker plot for the data set: 19, 22, 38, 44, 50, 54, 56, 62, 67, 79, 81, 85, 93.

Normal Distributions

Earlier in the text, you studied Gaussian models, which are examples of *normal distributions*. Recall that a **normal distribution** is modeled by a bell-shaped curve. This is called a **normal curve** and it is symmetric with respect to the mean.

Areas Under a Normal Curve

A normal distribution with mean $\bar{x}$ and standard deviation σ has the following properties:

• The total area under the normal curve is 1.

• About 68% of the area lies within 1 standard deviation of the mean.

• About 95% of the area lies within 2 standard deviations of the mean.

• About 99.7% of the area lies within 3 standard deviations of the mean.
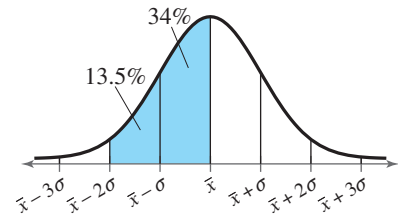
EXAMPLE 9   **Finding a Normal Probability**

A normal distribution has mean $\bar{x}$ and standard deviation σ. For a randomly selected x-value from the distribution, find the probability that $\bar{x} - 2\sigma \le x \le \bar{x}$.

Solution The probability that

$$\bar{x} - 2\sigma \le x \le \bar{x}$$

is the shaded area under the normal curve shown.

$$P(\bar{x} - 2\sigma \le x \le \bar{x}) \approx 0.135 + 0.34 = 0.475$$

✓ *Checkpoint* 🔊))) *Audio-video solution in English & Spanish at LarsonPrecalculus.com.*

A normal distribution has mean $\bar{x}$ and standard deviation σ. For a randomly selected x-value from the distribution, find the probability that $\bar{x} - \sigma \le x \le \bar{x} + 2\sigma$.

EXAMPLE 10   **Interpreting Normally Distributed Data**

The blood cholesterol readings for a group of women are normally distributed with a mean of 172 milligrams per deciliter and a standard deviation of 14 milligrams per deciliter.

a. About what percent of the women have readings between 158 and 186 milligrams per deciliter?

b. Readings less than 158 milligrams per deciliter are considered desirable. About what percent of the readings are desirable?

Solution

a. The readings of 158 and 186 milligrams per deciliter represent one standard deviation on either side of the mean. So, about 68% of the women have readings between 158 and 186 milligrams per deciliter.

b. A reading of 158 milligrams per deciliter is one standard deviation to the left of the mean. So, the percent of the readings that are desirable is about $0.15\% + 2.35\% + 13.5\% = 16\%$.

✓ *Checkpoint* 🔊))) *Audio-video solution in English & Spanish at LarsonPrecalculus.com.*

In Example 10, about what percent of the women have readings between 172 and 200 milligrams per deciliter? ■

The **standard normal distribution** is a normal distribution with mean 0 and standard deviation 1. Using the formula

$$z = \frac{x - \bar{x}}{\sigma}$$

transforms an x-value from a normal distribution with mean $\bar{x}$ and standard deviation σ into a corresponding z-value, or **z-score,** having a standard normal distribution. The z-score for an x-value is equal to the number of standard deviations the x-value lies above or below the mean $\bar{x}$. (See figure.)
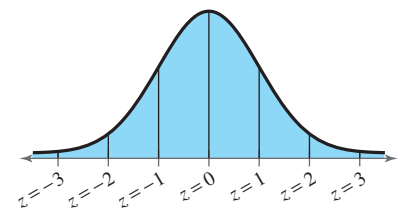
When z is a randomly selected value from a standard normal distribution, you can use the table on the next page to find the probability that z is less than or equal to some given value.

Standard Normal Table

z	.0	.1	.2	.3	.4	.5	.6	.7	.8	.9
−3	.0013	.0010	.0007	.0005	.0003	.0002	.0002	.0001	.0001	.0000+
−2	.0228	.0179	.0139	.0107	.0082	.0062	.0047	.0035	.0026	.0019
−1	.1587	.1357	.1151	.0968	.0808	.0668	.0548	.0446	.0359	.0287
−0	.5000	.4602	.4207	.3821	.3446	.3085	.2743	.2420	.2119	.1841
0	.5000	.5398	.5793	.6179	.6554	.6915	.7257	.7580	.7881	.8159
1	.8413	.8643	.8849	.9032	.9192	.9332	.9452	.9554	.9641	.9713
2	.9772	.9821	.9861	.9893	.9918	.9938	.9953	.9965	.9974	.9981
3	.9987	.9990	.9993	.9995	.9997	.9998	.9998	.9999	.9999	1.0000–

For example, to find the value of $P(z \leq -0.4)$, find the value in the table where the row labeled "−0" and the column labeled ".4" intersect. The table shows that

$$P(z \leq -0.4) = 0.3446.$$

You can also use the standard normal table to find probabilities for normal distributions by first converting values from the distribution to z-scores. Example 11 illustrates this.

EXAMPLE 11 Using the Standard Normal Table

Scientists conducted aerial surveys of a seal sanctuary and recorded the number x of seals they observed during each survey. The numbers of seals observed were normally distributed with a mean of 73 and a standard deviation of 14.1. Find the probability that the scientists observed at most 50 seals during a survey.

Solution Begin by finding the z-score that corresponds to an x-value of 50.

$$z = \frac{x - \bar{x}}{\sigma} = \frac{50 - 73}{14.1} \approx -1.6$$

Then use the table to find $P(x \leq 50) \approx P(z \leq -1.6) = 0.0548$. So, the probability that the scientists observed at most 50 seals during a survey is about 0.0548.

Example 11 uses a z-score and the standard normal table to find the probability of observing seals at a sanctuary.

✓ Checkpoint ◀))) *Audio-video solution in English & Spanish at LarsonPrecalculus.com.*

In Example 11, find the probability that the scientists observed at most 90 seals during a survey.

Summarize (Section B.2)

1. State the definitions of the mean, median, and mode of a set of data *(page B8)*. For an example of comparing measures of central tendency, see Example 1.

2. Explain how to determine the measure of central tendency that best represents a set of data *(page B9)*. For an example of determining the measures of central tendency that best represent sets of data, see Example 2.

3. Explain how to find the standard deviation of a set of data *(pages B10 and B11)*. For examples of finding the standard deviations of sets of data, see Examples 4–6.

4. Explain how to sketch a box-and-whisker plot *(page B13)*. For an example of sketching box-and-whisker plots, see Example 8.

5. Describe what is meant by normally distributed data *(pages B14–B16)*. For examples of interpreting normally distributed data, see Examples 9–11.

B.2   Exercises

Vocabulary: Fill in the blanks.

1. The _____ of n numbers is the sum of the numbers divided by n.

2. If there is an even number of data values in a data set, then the _____ is the average of the two middle numbers.

3. When two numbers of a data set tie for most frequent occurrence, the collection has two _____ and is called _____.

4. Two measures of dispersion associated with the mean are called the _____ and the _____ _____ of a data set.

5. _____ measure dispersion associated with the median.

6. The _____ _____ of a data set is the difference of the upper quartile and the lower quartile.

7. A value that is widely separated from the rest of the data in a data set is called an _____.

8. You can represent quartiles graphically by creating a _____ _____.

9. The _____ _____ distribution is a normal distribution with mean 0 and standard deviation 1.

10. The _____ for an x-value is equal to the number of standard deviations the x-value lies above or below the mean in a standard normal distribution.

Skills and Applications

Finding Measures of Central Tendency In Exercises 11–16, find the mean, median, and mode of the data set.

11. 5, 12, 7, 14, 8, 9, 7

12. 30, 37, 32, 39, 33, 34, 32

13. 5, 12, 7, 24, 8, 9, 7

14. 20, 37, 32, 39, 33, 34, 32

15. 5, 12, 7, 14, 9, 7

16. 30, 37, 32, 39, 34, 32

17. **Electric Bills** A household had the following monthly bills for electricity. What are the mean and median of the collection of bills? *(Spreadsheet at LarsonPrecalculus.com)*

January	$67.92	February	$59.84
March	$52.00	April	$52.50
May	$57.99	June	$65.35
July	$81.76	August	$74.98
September	$87.82	October	$83.18
November	$65.35	December	$57.00

18. **Car Rental** A car rental company kept the following record of the numbers of rental cars driven. What are the mean, median, and mode of the data? *(Spreadsheet at LarsonPrecalculus.com)*

Monday	410	Tuesday	260
Wednesday	320	Thursday	320
Friday	460	Saturday	150

19. **Families** Researchers conducted a study on families with six children. The table shows the numbers of families in the study with the indicated numbers of girls. Determine the mean, median, and mode of this set of data.

Number of Girls	0	1	2	3	4	5	6
Frequency	1	24	45	54	50	19	7

    Spreadsheet at LarsonPrecalculus.com

20. **Sports** A baseball fan examined the records of a favorite baseball player's performance during his last 50 games. The table shows the numbers of games in which the player had 0, 1, 2, 3, and 4 hits.

Number of Hits	0	1	2	3	4
Frequency	14	26	7	2	1

    (a) Determine the mean number of hits per game.

    (b) Determine the player's batting average, assuming he had 200 at-bats during the 50-game series.

21. **Test Scores** A professor records the following scores for a 100-point exam. *(Spreadsheet at LarsonPrecalculus.com)*

    99, 64, 80, 77, 59,
    72, 87, 79, 92, 88,
    90, 42, 20, 89, 42,
    100, 98, 84, 78, 91

    Which measure of central tendency best describes these test scores?

22. Shoe Sales A salesman sold eight pairs of men's black dress shoes. The sizes of the eight pairs were as follows: $10\frac{1}{2}$, 8, 12, $10\frac{1}{2}$, 10, $9\frac{1}{2}$, 11, and $10\frac{1}{2}$. Which measure (or measures) of central tendency best describes the typical shoe size for the data? *(Spreadsheet at LarsonPrecalculus.com)*

23. Think About It The histograms represent the test scores of two classes of a college course in mathematics. Which histogram shows the lesser standard deviation? Explain your reasoning.
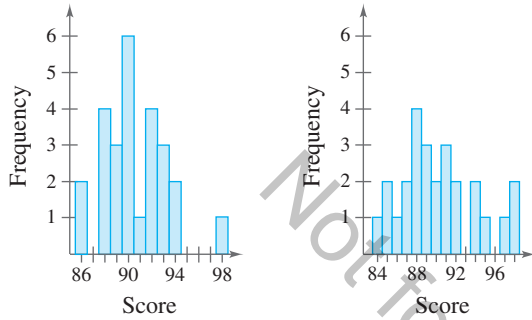
24. Test Scores An instructor adds five points to each student's exam score. Does this change the mean or standard deviation of the exam scores? Explain.

Finding Mean and Standard Deviation In Exercises 25 and 26, line plots of sets of data are given. Find the mean and standard deviation of each set of data.

25. (a)

(b)

26. (a)

(b)
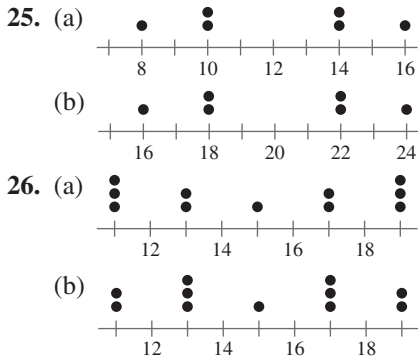
Finding Mean, Variance, and Standard Deviation In Exercises 27–32, find the mean ($\bar{x}$), variance (v), and standard deviation (σ) of the data set.

27. 4, 10, 8, 2
28. 3, 15, 6, 9, 2
29. 0, 1, 1, 2, 2, 2, 3, 3, 4
30. 2, 2, 2, 2, 2, 2
31. 49, 62, 40, 29, 32, 70
32. 1.5, 0.4, 2.1, 0.7, 0.8

Using the Alternative Formula for Standard Deviation In Exercises 33–38, use the alternative formula for standard deviation to find the standard deviation of the data set.

33. 2, 4, 6, 6, 13, 5
34. 10, 25, 50, 26, 15, 33, 29, 4
35. 246, 336, 473, 167, 219, 359

36. 6.0, 9.1, 4.4, 8.7, 10.4
37. 8.1, 6.9, 3.7, 4.2, 6.1
38. 9.0, 7.5, 3.3, 7.4, 6.0

39. Price of Gold The following data represent the average prices of gold (in dollars per troy ounce) for the years 1992 through 2011. Use a software program or a graphing utility to find the mean, variance, and standard deviation of the data. What percent of the data lies within two standard deviations of the mean? *(Source: U.S. Bureau of Mines and U.S. Geological Survey)* *(Spreadsheet at LarsonPrecalculus.com)*

345, 361, 385, 386, 389, 332, 295, 280, 280, 272, 311, 365, 411, 446, 606, 699, 874, 975, 1227, 1600

40. Quartiles and Box-and-Whisker Plots Find the lower and upper quartiles and sketch a box-and-whisker plot for each data set.

(a) 11, 10, 11, 14, 17, 16, 14, 11, 8, 14, 20
(b) 46, 48, 48, 50, 52, 47, 51, 47, 49, 53
(c) 19, 12, 14, 9, 14, 15, 17, 13, 19, 11, 10, 19
(d) 20.1, 43.4, 34.9, 23.9, 33.5, 24.1, 22.5, 42.4, 25.7, 17.4, 23.8, 33.3, 17.3, 36.4, 21.8

Product Lifetime

41. A manufacturer has redesigned a machine part in an attempt to increase the lifetime of the part. The two sets of data list the lifetimes (in months) of 20 units with the original design and 20 units with the new design. Construct a box-and-whisker plot for each set of data. *(Spreadsheet at LarsonPrecalculus.com)*

Original Design

15.1	78.3	56.3	68.9	30.6
27.2	12.5	42.7	72.7	20.2
53.0	13.5	11.0	18.4	85.2
10.8	38.3	85.1	10.0	12.6

New Design

55.8	71.5	25.6	19.0	23.1
37.2	60.0	35.3	18.9	80.5
46.7	31.1	67.9	23.5	99.5
54.0	23.2	45.5	24.8	87.8

42. Explain the differences between the box-and-whisker plots you constructed in Exercise 41.

Finding a Normal Probability   In Exercises 43–46, a normal distribution has mean $\bar{x}$ and standard deviation σ. Find the indicated probability for a randomly selected x-value from the distribution.

43. $P(x \leq \bar{x} - \sigma)$

44. $P(x \geq \bar{x} + 2\sigma)$

45. $P(\bar{x} - \sigma \leq x \leq \bar{x} + \sigma)$

46. $P(\bar{x} - 3\sigma \leq x \leq \bar{x})$

Normal Distribution   In Exercises 47–50, a normal distribution has a mean of 33 and a standard deviation of 4. Find the probability that a randomly selected x-value from the distribution is in the given interval.

47. Between 29 and 37

48. Between 33 and 45

49. At least 29

50. At most 37

Using the Standard Normal Table   In Exercises 51–56, a normal distribution has a mean of 64 and a standard deviation of 7. Use the standard normal table to find the indicated probability for a randomly selected x-value from the distribution.

51. $P(x \leq 68)$

52. $P(x \leq 80)$

53. $P(x \geq 45)$

54. $P(x \geq 75)$

55. $P(60 \leq x \leq 75)$

56. $P(45 \leq x \leq 65)$

57. Biology   A study found that the wing lengths of houseflies are normally distributed with a mean of about 4.6 millimeters and a standard deviation of about 0.4 millimeter. What is the probability that a randomly selected housefly has a wing length of at least 5 millimeters?

58. Boxes of Cereal   A machine fills boxes of cereal. The weights of cereal in the boxes are normally distributed with a mean of 20 ounces and a standard deviation of 0.25 ounce.

  (a) Find the z-scores for weights of 19.4 ounces and 20.4 ounces.

  (b) What is the probability that a randomly selected cereal box weighs at most 19.4 ounces?

  (c) What is the probability that a randomly selected cereal box weighs between 19.4 and 20.4 ounces?

Exploration

True or False?   In Exercises 59–62, determine whether the statement is true or false. Justify your answer.

59. There are some quantitative data sets that do not have medians.

60. About one quarter of the data in a set is less than the lower quartile.

61. To construct a box-and-whisker plot for a data set, you need to know the least number, the lower quartile, the mean, the upper quartile, and the greatest number of the data set.

62. It is not possible for a value from a normal distribution to have a z-score equal to 0.

63. Writing   When $n\%$ of the values in a data set are less than or equal to a certain value, that value is called the **nth percentile.** For normally distributed data, describe the value that represents the 84th percentile in terms of the mean and standard deviation.

64.   **HOW DO YOU SEE IT?**   The fuel efficiencies of a brand of automobile are normally distributed with a mean of 26 miles per gallon and a standard deviation of 2.4 miles per gallon, as shown in the figure.
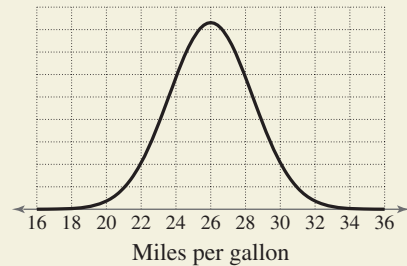
Miles per gallon

  (a) Which is greater, the probability of choosing a car at random that gets between 26 and 28 miles per gallon or the probability of choosing a car at random that gets between 22 and 24 miles per gallon?

  (b) Which is greater, the probability of choosing a car at random that gets between 20 and 22 miles per gallon or the probability of choosing a car at random that gets at least 30 miles per gallon?

65. Reasoning   Compare your answers for Exercises 11 and 13 and those for Exercises 12 and 14. Which of the measures of central tendency is affected by outliers? Explain your reasoning.

66. Reasoning

  (a) Add 6 to each measurement in Exercise 11 and calculate the mean, median, and mode of the revised measurements. How do the measures of central tendency change?

  (b) When you add a constant k to each measurement in a set of data, how do the measures of central tendency change?

67. Reasoning   Without calculating the standard deviation, explain why the data set $\{4, 4, 20, 20\}$ has a standard deviation of 8.

68. Think About It   Construct a collection of numbers that has the following measures of central tendency. If this is not possible, then explain why.

Mean $= 6$, median $= 4$, mode $= 4$

B.3   Modeling Data

■ Use the correlation coefficient to measure how well a model fits a set of data.
■ Use the sum of square differences to measure how well a model fits a set of data.
■ Find the least squares regression line for a set of data.
■ Find the least squares regression parabola for a set of data.
■ Describe misuses of statistics.

Throughout this text, you have been using or finding models for data sets. For instance, in a previous section, you determined how closely a given model represents a data set, and you used the *regression* feature of a graphing utility to find a linear model for a data set. This section expands upon the concepts learned regarding the use of least squares regression to fit models to data.

The method of least squares provides a way of creating mathematical models for sets of data. For instance, in Exercise 19 on page B25, you will find the least squares regression line for the average prices of generic prescription drugs in the United States from 2007 through 2010.
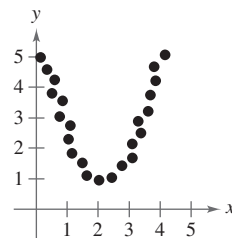
Correlation

A **correlation** can be defined as a relationship between two variables. A scatter plot can help you determine whether a correlation exists between two variables. The scatter plots below illustrate several types of correlation.
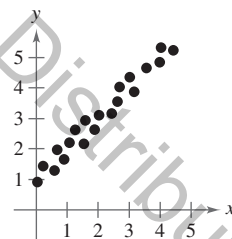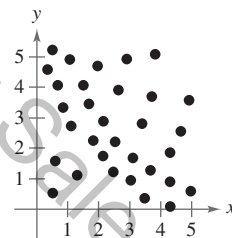

Positive linear correlation: y tends to increase as x increases.
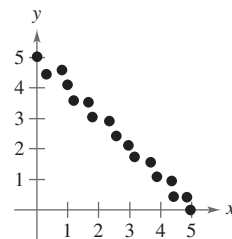
Negative linear correlation: y tends to decrease as x increases.

Nonlinear correlation

No correlation

Recall that the **correlation coefficient** r gives a measure of how well a model fits a set of data. The range of r is $[-1, 1]$. For a linear model, r close to 1 indicates a strong positive linear correlation between x and y, r close to -1 indicates a strong negative linear correlation, and r close to 0 indicates a weak or no *linear* correlation. A formula for r is

▷ **ALGEBRA HELP**   Note that the formula for r uses summation notation.

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

where n is the number of data points (x, y).

Sum of Square Differences

The *regression* feature of a graphing utility uses the **method of least squares** to find a mathematical model for a set of data. As a measure of how well a model fits a set of data points

$$\{(x_1, y_1), (x_2, y_2), . . . , (x_n, y_n)\}$$

you can add the squares of the differences between the actual y-values and the values given by the model to obtain the **sum of square differences.**

EXAMPLE 1 Finding the Sum of Square Differences

The table shows the heights x (in feet) and the diameters y (in inches) of eight trees. Find the sum of square differences for the linear model $y* = 0.54x - 29.5$.

DATA	x	70	72	75	76	85	78	77	80
	y	8.3	10.5	11.0	11.4	12.9	14.0	16.3	18.0

Spreadsheet at LarsonPrecalculus.com

Solution Evaluate $y* = 0.54x - 29.5$ for each given value of x. Then find the difference $y - y*$ using each value of y and the corresponding value of $y*$, and square the difference.

x	70	72	75	76	85	78	77	80
y	8.3	10.5	11.0	11.4	12.9	14.0	16.3	18.0
$y*$	8.3	9.38	11.0	11.54	16.4	12.62	12.08	13.7
$y - y*$	0	1.12	0	-0.14	-3.5	1.38	4.22	4.3
$(y - y*)^2$	0	1.2544	0	0.0196	12.25	1.9044	17.8084	18.49

Finally, add the values in the last row to find the sum of square differences.

$$0 + 1.2544 + 0 + 0.0196 + 12.25 + 1.9044 + 17.8084 + 18.49 = 51.7268$$

✓ **Checkpoint** ◀))) *Audio-video solution in English & Spanish at LarsonPrecalculus.com.*

The table shows the shoe sizes x and the heights y (in inches) of eight men. Find the sum of square differences for the linear model $y* = 1.87x + 51.2$.

DATA	x	8.5	9.0	9.0	9.5	10.0	10.5	11.0	12.0
	y	66.0	68.5	67.5	70.0	72.0	69.5	71.5	73.5

Spreadsheet at LarsonPrecalculus.com ■

Least Squares Regression Lines

The linear model that has the least sum of squared differences is the **least squares regression line** for the data and is the best-fitting linear model for the data. To find the least squares regression line $y = ax + b$ for the points $\{(x_1, y_1), (x_2, y_2), . . . , (x_n, y_n)\}$ algebraically, you need to solve the following system for a and b.

•• **REMARK** Deriving this system requires the use of *partial derivatives*, which you will study in a calculus course.

$$\begin{cases} nb + \left(\sum_{i=1}^{n} x_i \right)a = \sum_{i=1}^{n} y_i \\ \left(\sum_{i=1}^{n} x_i \right)b + \left(\sum_{i=1}^{n} x_i^2 \right)a = \sum_{i=1}^{n} x_i y_i \end{cases}$$

<div style="background:#c00;color:white;">EXAMPLE 2</div>   **Finding a Least Squares Regression Line**

Find the least squares regression line for the points $(-3, 0)$, $(-1, 1)$, $(0, 2)$, and $(2, 3)$.

Solution   Construct the following table to help you come up with the appropriate system of equations.

:::::::::::::::: ▷

:: **REMARK**   The least squares regression line for the data given in Example 1 is $y \approx 0.43x - 20.3$ and the sum of square differences is about 43.3. Try verifying this.
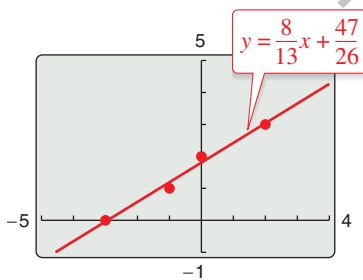
x	y	xy	x^2
-3	0	0	9
-1	1	-1	1
0	2	0	0
2	3	6	4
$\displaystyle\sum_{i=1}^{n} x_i = -2$	$\displaystyle\sum_{i=1}^{n} y_i = 6$	$\displaystyle\sum_{i=1}^{n} x_i y_i = 5$	$\displaystyle\sum_{i=1}^{n} x_i^2 = 14$

$y = \dfrac{8}{13}x + \dfrac{47}{26}$

Figure B.3

Applying the system for the least squares regression line with $n = 4$ produces

$$\begin{cases} nb + \left(\displaystyle\sum_{i=1}^{n} x_i\right)a = \displaystyle\sum_{i=1}^{n} y_i \\ \left(\displaystyle\sum_{i=1}^{n} x_i\right)b + \left(\displaystyle\sum_{i=1}^{n} x_i^2\right)a = \displaystyle\sum_{i=1}^{n} x_i y_i \end{cases} \implies \begin{cases} 4b - 2a = 6 \\ -2b + 14a = 5 \end{cases}.$$

Solving this system of equations produces

$$a = \frac{8}{13} \quad \text{and} \quad b = \frac{47}{26}.$$

So, the least squares regression line is $y = \frac{8}{13}x + \frac{47}{26}$, as shown in Figure B.3.

✓ **Checkpoint** 🔊)) *Audio-video solution in English & Spanish at LarsonPrecalculus.com.*

Find the least squares regression line for the points $(-1, -1)$, $(0, 0)$, $(1, 2)$, and $(2, 4)$.

<div style="background:#c00;color:white;">EXAMPLE 3</div>   **Finding a Correlation Coefficient**

Find the correlation coefficient r for the data given in Example 2. How well does the linear model fit the data?

Solution   Using the formula given on page B20 with $n = 4$,

Least squares regression analysis has a wide variety of real-life applications. For instance, in Exercise 21, you will find the least squares regression line that models the demand for a tool at a hardware retailer as a function of price.

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

$$= \frac{4(5) - (-2)(6)}{\sqrt{4(14) - (-2)^2}\sqrt{4(14) - (6)^2}}$$

$$\approx 0.992.$$

Because r is close to 1, there is a strong positive linear correlation between x and y. So, the least squares regression line found in Example 2 fits the data very well.

✓ **Checkpoint** 🔊)) *Audio-video solution in English & Spanish at LarsonPrecalculus.com.*

Find the correlation coefficient r for the data given in the Checkpoint with Example 2. How well does the linear model fit the data?   ■

mihalec/Shutterstock.com

Least Squares Regression Parabolas

The **least squares regression parabola** $y = ax^2 + bx + c$ for the points

$$\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$$

is obtained by solving the following system for a, b, and c.

$$\begin{cases} nc + \left(\sum_{i=1}^{n} x_i\right)b + \left(\sum_{i=1}^{n} x_i^2\right)a = \sum_{i=1}^{n} y_i \\ \left(\sum_{i=1}^{n} x_i\right)c + \left(\sum_{i=1}^{n} x_i^2\right)b + \left(\sum_{i=1}^{n} x_i^3\right)a = \sum_{i=1}^{n} x_i y_i \\ \left(\sum_{i=1}^{n} x_i^2\right)c + \left(\sum_{i=1}^{n} x_i^3\right)b + \left(\sum_{i=1}^{n} x_i^4\right)a = \sum_{i=1}^{n} x_i^2 y_i \end{cases}$$

EXAMPLE 4 **Finding a Least Squares Regression Parabola**

Find the least squares regression parabola for the points $(1, 2)$, $(2, 1)$, $(3, 2)$, and $(4, 4)$.

Solution Construct the following table to help you come up with the appropriate system of equations.

x	x^2	x^3	x^4	y	xy	x^2y
1	1	1	1	2	2	2
2	4	8	16	1	2	4
3	9	27	81	2	6	18
4	16	64	256	4	16	64

From the table, the sums you need are as follows.

$$\sum_{i=1}^{n} x_i = 10, \qquad \sum_{i=1}^{n} x_i^2 = 30, \qquad \sum_{i=1}^{n} x_i^3 = 100, \qquad \sum_{i=1}^{n} x_i^4 = 354,$$

$$\sum_{i=1}^{n} y_i = 9, \qquad \sum_{i=1}^{n} x_i y_i = 26, \qquad \text{and} \qquad \sum_{i=1}^{n} x_i^2 y_i = 88$$

Applying the system for the least squares regression parabola with $n = 4$ produces

$$\begin{cases} nc + \left(\sum_{i=1}^{n} x_i\right)b + \left(\sum_{i=1}^{n} x_i^2\right)a = \sum_{i=1}^{n} y_i \\ \left(\sum_{i=1}^{n} x_i\right)c + \left(\sum_{i=1}^{n} x_i^2\right)b + \left(\sum_{i=1}^{n} x_i^3\right)a = \sum_{i=1}^{n} x_i y_i \\ \left(\sum_{i=1}^{n} x_i^2\right)c + \left(\sum_{i=1}^{n} x_i^3\right)b + \left(\sum_{i=1}^{n} x_i^4\right)a = \sum_{i=1}^{n} x_i^2 y_i \end{cases} \implies \begin{cases} 4c + 10b + 30a = 9 \\ 10c + 30b + 100a = 26. \\ 30c + 100b + 354a = 88 \end{cases}$$

Solving this system of equations produces $a = \frac{3}{4}$, $b = -\frac{61}{20}$, and $c = \frac{17}{4}$. So, the least squares regression parabola is

$$y = \frac{3}{4}x^2 - \frac{61}{20}x + \frac{17}{4}$$
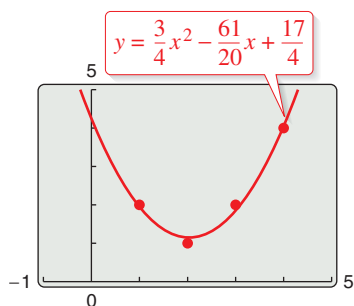
as shown in Figure B.4.

$y = \frac{3}{4}x^2 - \frac{61}{20}x + \frac{17}{4}$

Figure B.4

✓ *Checkpoint* 🔊)) *Audio-video solution in English & Spanish at LarsonPrecalculus.com.*

Find the least squares regression parabola for the points $(-3, -2)$, $(-2, 0)$, $(-1, 1)$, and $(1, 0)$.

Misuses of Statistics

One way of misusing statistics is to use a misleading graph of the data. Such graphs can lead to false conclusions. For instance, the following graph is misleading because the break in the vertical axis gives the impression that first quarter sales were disproportionately more than those of other quarters.
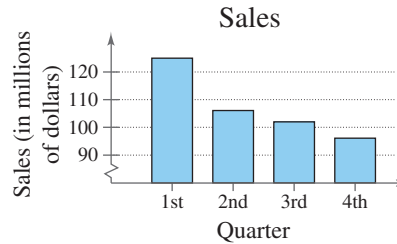
A way to redraw this graph so it is not misleading would be to show it without the break, as follows. Notice that in the redrawn graph, the difference in first quarter sales and those of other quarters does not appear to be as dramatic.
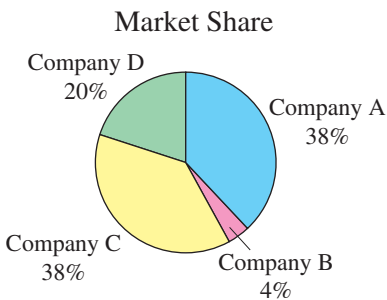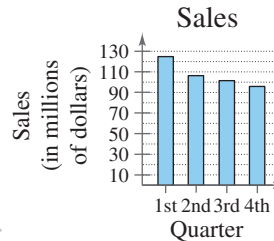
To illustrate further, consider the graph at the right. This graph is misleading because the angle at which it is shown gives the impression that Company C has a greater market share than the other companies. Redrawing the graph as shown in Figure B.5 clarifies the fact that Company A and Company C have equal market shares.
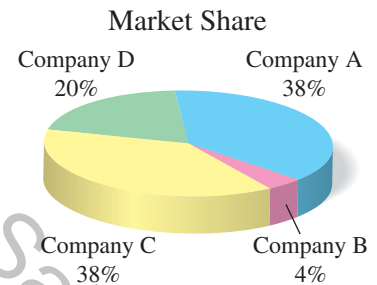
Figure B.5

Summarize (Section B.3)

1. Explain how to use the correlation coefficient to measure how well a model fits a set of data *(page B20)*. For an example of finding a correlation coefficient, see Example 3.

2. Explain how to use the sum of square differences to measure how well a model fits a set of data *(page B21)*. For an example of finding the sum of square differences for a linear model, see Example 1.

3. Describe the steps involved in finding the least squares regression line for a set of data *(page B21)*. For an example of finding the least squares regression line for a set of data, see Example 2.

4. Describe the steps involved in finding the least squares regression parabola for a set of data *(page B23)*. For an example of finding the least squares regression parabola for a set of data, see Example 4.

5. Give examples of misuses of statistics *(page B24)*.

B.3   Exercises

Vocabulary: **Fill in the blanks.**

1. A _____ can be defined as a relationship between two variables.
2. A graphing utility uses the _____ of _____ _____ to find a mathematical model for a set of data.
3. The _____ coefficient and the sum of _____ _____ measure how well a model fits a set of data points.
4. The _____ _____ _____ line for a set of data is the linear model that has the least sum of squared differences.

Skills and Applications

Finding the Sum of Square Differences In Exercises 5–12, find the sum of square differences for the set of data points and the model.

5. $(-3, -1), (-1, 0), (0, 2), (2, 3), (4, 4)$
   $y = 0.5x + 0.5$
6. $(0, 2), (1, 1), (2, 2), (3, 4), (5, 6)$
   $y = 0.8x + 2$
7. $(-2, 6), (-1, 4), (0, 2), (1, 1), (2, 1)$
   $y = -1.7x + 2.7$
8. $(0, 7), (2, 5), (3, 2), (4, 3), (6, 0)$
   $y = -1.2x + 7$
9. $(-3, 2), (-2, 2), (-1, 4), (0, 6), (1, 8)$
   $y = 0.29x^2 + 2.2x + 6$
10. $(-3, 4), (-1, 2), (1, 1), (3, 0)$
    $y = 0.06x^2 - 0.7x + 1$
11. $(0, 10), (1, 9), (2, 6), (3, 0)$
    $y = -1.25x^2 + 0.5x + 10$
12. $(-1, -4), (1, -3), (2, 0), (4, 5), (6, 9)$
    $y = 0.14x^2 + 1.3x - 3$

Finding a Least Squares Regression Line In Exercises 13–16, find the least squares regression line for the points. Verify your answer with a graphing utility.

13. $(-4, 1), (-3, 3), (-2, 4), (-1, 6)$
14. $(0, -1), (2, 0), (4, 3), (6, 5)$
15. $(-3, 1), (-1, 2), (1, 2), (4, 3)$
16. $(0, -1), (2, 1), (3, 2), (5, 3)$

17. **Finding Correlation Coefficients** Find the correlation coefficient r for the data given in (a) Exercise 13 and (b) Exercise 15. How well does each linear model fit the corresponding data?
18. **Finding Correlation Coefficients** Find the correlation coefficient r for the data given in (a) Exercise 14 and (b) Exercise 16. How well does each linear model fit the corresponding data?

19. **Prescription Prices**
    The average prices of generic prescription drugs in the United States from 2007 through 2010 are represented by the ordered pairs (x, y), where x represents the year, with $x = 7$ corresponding to 2007 and y represents the average price (in dollars). Find the least squares regression line for the data. What is the sum of square differences? *(Source: National Association of Chain Drug Stores)*

    $(7, 32.60), (8, 35.21), (9, 39.25), (10, 44.14)$

20. **College Enrollment** The projected enrollments in U.S. private colleges from 2017 through 2020 are represented by the ordered pairs (x, y), where x represents the year, with $x = 17$ corresponding to 2017 and y represents the projected enrollment (in millions of students). Find the least squares regression line for the data. What is the sum of square differences? *(Source: U.S. National Center for Education Statistics)*

    $(17, 6.137), (18, 6.217), (19, 6.290), (20, 6.340)$

21. **Demand** A hardware retailer wants to know the demand y for a tool as a function of price x. The table lists the monthly sales for four different prices of the tool.

Price, x	$25	$30	$35	$40
Demand, y	82	75	67	55

    (a) Find the least squares regression line for the data.
    (b) Estimate the demand when the price is $32.95.
   (c) What price will create a demand of 83 tools?

22. Agriculture An agronomist used four test plots to determine the relationship between the wheat yield y (in bushels per acre) and the amount of fertilizer x (in pounds per acre). The table shows the results.

| Fertilizer, x	100	150	200	250
Yield, y	35	44	50	56

(a) Find the least squares regression line for the data.

(b) Estimate the yield for a fertilizer application of 160 pounds per acre.

Finding a Least Squares Regression Parabola
In Exercises 23–30, find the least squares regression parabola for the points. Verify your answer with a graphing utility.

23. $(0, 0)$, $(2, 4)$, $(4, 2)$

24. $(-2, 6)$, $(-1, 2)$, $(1, 3)$

25. $(-1, 4)$, $(0, 2)$, $(1, 0)$, $(3, 4)$

26. $(-3, -1)$, $(-1, 2)$, $(1, 2)$, $(3, 0)$

27. $(-2, 18)$, $(-1, 9)$, $(0, 4)$, $(1, 3)$, $(2, 6)$

28. $(0, 0)$, $(1, -0.75)$, $(4, -21)$, $(5, -33)$, $(7, -68)$

29.

30.
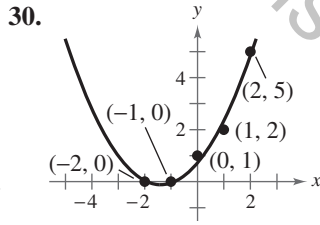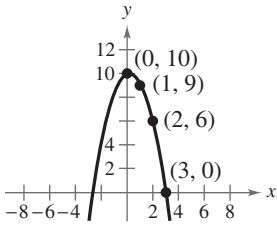
Misleading Graph **In Exercises 31 and 32, explain why the graph could be misleading. Then redraw the graph so it is not misleading.**

31.

32.
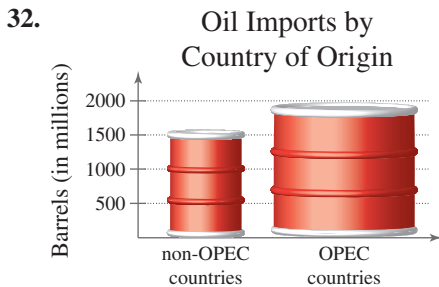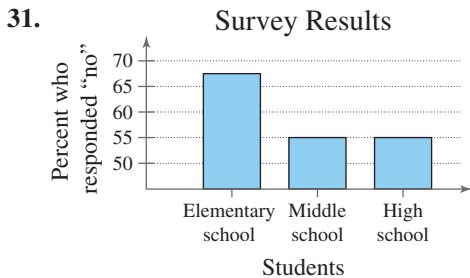
Exploration

True or False? **In Exercises 33–36, determine whether the statement is true or false. Justify your answer.**

33. Data that are modeled by $y = -0.238x + 25$ have a negative linear correlation.

34. When the correlation coefficient is $r \approx -0.98781$, the model is a good fit.

35. A correlation coefficient of $r \approx 0.021$ implies that the data have weak or no linear correlation.

36. A linear regression model with a positive correlation coefficient will have a slope that is greater than 0.

37. Writing A linear model for predicting prize winnings at a race is based on data for 3 years. Write a paragraph discussing the potential accuracy or inaccuracy of such a model.

38. HOW DO YOU SEE IT? Match the scatter plot with the correct correlation coefficient. Explain your reasoning. [The scatter plots are labeled (i), (ii), (iii), (iv), (v), and (vi).]
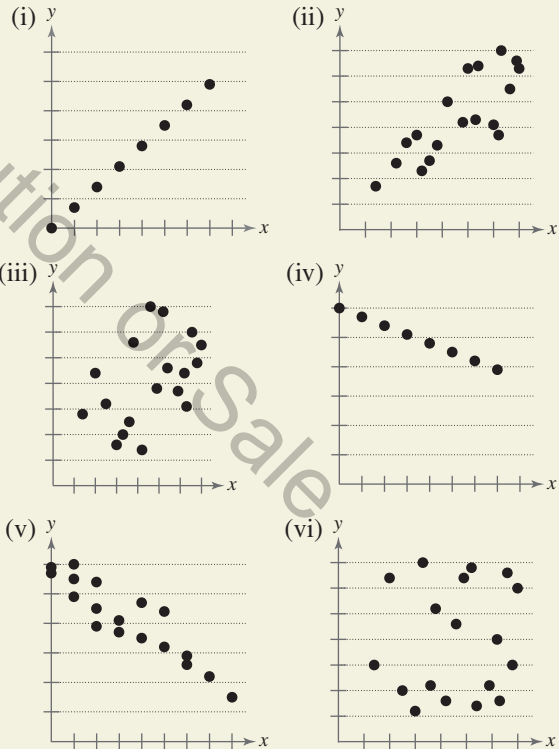
(a) $r = -1$ (b) $r = 1$ (c) $r = 0.04$

(d) $r = 0.45$ (e) $r = -0.92$ (f) $r = 0.81$

39. Research Project Use your school's library, the Internet, or some other reference source to find examples of real-life data (other than those discussed in this chapter) that can be modeled using a least squares regression line or a least squares regression parabola.

Chapter Summary

	What Did You Learn?	**Explanation/Examples**	**Review Exercises**
Section B.1	Understand terminology associated with statistics *(p. B1)*, use line plots to order and analyze data *(p. B3)*, use stem-and-leaf plots to organize and compare data *(p. B4)*, and use histograms to represent frequency distributions *(p. B5)*.	Data consist of information that comes from observations, responses, counts, or measurements. Samples are subsets, or parts, of a population. Types of sampling methods include random, stratified random, systematic, convenience, and self-selected. See pages B1 and B2 for more terminology.	1–4

Line Plot

High temperatures

Stem-and-Leaf Plot

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

Key: $5|8 = 58$

Frequency Distribution

Interval	Tally
[7, 9)	|
[9, 11)	||||
[11, 13)	ﾱ ﾱ ﾱ
[13, 15)	ﾱ ﾱ ﾱ ﾱ ﾱ ||
[15, 17)	|||
[17, 19)	|

Histogram

Percent of population 65 or older

	What Did You Learn?	**Explanation/Examples**	**Review Exercises**
Section B.2	Find and interpret the mean, median, and mode of a set of data *(p. B8)*, and determine the measure of central tendency that best represents a set of data *(p. B9)*.	The mean of *n* numbers is the sum of the numbers divided by *n*. The median of *n* numbers is the middle number when the numbers are written in order. When *n* is even, the median is the average of the two middle numbers. The mode of *n* numbers is the number that occurs most frequently.	5–10
	Find the standard deviation of a set of data *(p. B10)*.	The variance of a set of numbers $\{x_1, x_2, \ldots, x_n\}$ with a mean of $\bar{x}$ is $$v = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n}$$ and the standard deviation is $\sigma = \sqrt{v}$. **Alternative Formula for Standard Deviation** $$\sigma = \sqrt{\frac{x_1^2 + x_2^2 + \cdots + x_n^2}{n} - \bar{x}^2}$$	11–16
	Create and use box-and-whisker plots *(p. B13)*.	**Box-and-Whisker Plot** 	17, 18

	What Did You Learn?	**Explanation/Examples**	**Review Exercises**
Section B.2	Interpret normally distributed data *(p. B14)*.	Normal distribution with mean $\bar{x}$ and standard deviation σ: Using the formula $z = (x - \bar{x})/\sigma$ transforms an x-value into a corresponding z-score having a standard normal distribution.	19–28
Section B.3	Use the correlation coefficient to measure how well a model fits a set of data *(p. B20)*.	The correlation coefficient r gives a measure of how well a model fits a set of data. A formula for r is $$r = \frac{n\sum\limits_{i=1}^{n}x_iy_i - \left(\sum\limits_{i=1}^{n}x_i\right)\left(\sum\limits_{i=1}^{n}y_i\right)}{\sqrt{n\sum\limits_{i=1}^{n}x_i^2 - \left(\sum\limits_{i=1}^{n}x_i\right)^2}\sqrt{n\sum\limits_{i=1}^{n}y_i^2 - \left(\sum\limits_{i=1}^{n}y_i\right)^2}}$$ where n is the number of data points (x, y). For a linear model, r close to 1 indicates a strong positive linear correlation between x and y, r close to -1 indicates a strong negative linear correlation, and r close to 0 indicates a weak or no *linear* correlation.	29, 30
	Use the sum of square differences to measure how well a model fits a set of data *(p. B21)*.	As a measure of how well a model fits a set of data points, add the squares of the differences between the actual y-values and the values given by the model to obtain the sum of square differences.	31, 32
	Find the least squares regression line *(p. B21)* and the least squares regression parabola *(p. B23)* for a set of data.	To find the least squares regression line $y = ax + b$ for the points $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ algebraically, solve the following system for a and b. $$\begin{cases} nb + \left(\sum\limits_{i=1}^{n}x_i\right)a = \sum\limits_{i=1}^{n}y_i \\ \left(\sum\limits_{i=1}^{n}x_i\right)b + \left(\sum\limits_{i=1}^{n}x_i^2\right)a = \sum\limits_{i=1}^{n}x_iy_i \end{cases}$$ To find the least squares regression parabola $y = ax^2 + bx + c$ for the points $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ algebraically, solve the following system for a, b, and c. $$\begin{cases} nc + \left(\sum\limits_{i=1}^{n}x_i\right)b + \left(\sum\limits_{i=1}^{n}x_i^2\right)a = \sum\limits_{i=1}^{n}y_i \\ \left(\sum\limits_{i=1}^{n}x_i\right)c + \left(\sum\limits_{i=1}^{n}x_i^2\right)b + \left(\sum\limits_{i=1}^{n}x_i^3\right)a = \sum\limits_{i=1}^{n}x_iy_i \\ \left(\sum\limits_{i=1}^{n}x_i^2\right)c + \left(\sum\limits_{i=1}^{n}x_i^3\right)b + \left(\sum\limits_{i=1}^{n}x_i^4\right)a = \sum\limits_{i=1}^{n}x_i^2y_i \end{cases}$$	33–36
	Describe misuses of statistics *(p. B24)*.	One way of misusing statistics is to use a misleading graph of the data. Such graphs can lead to false conclusions. See page B24 for examples.	37, 38

Review Exercises

B.1 **Constructing a Line Plot** **In Exercises 1 and 2, construct a line plot of the data.**

1. 11, 14, 21, 17, 13, 13, 17, 17, 18, 14, 21, 11, 14, 15

2. 6, 8, 7, 9, 4, 4, 6, 8, 8, 8, 9, 5, 5, 9, 9, 4, 2, 2, 6, 8

3. **Home Runs** The numbers of home runs hit by the 20 baseball players with the best single-season batting averages in Major League Baseball since 1900 are shown below. Construct a stem-and-leaf plot of the data. *(Source: Major League Baseball) (Spreadsheet at LarsonPrecalculus.com)*

   DATA 14, 25, 8, 8, 7, 7, 19, 37, 39, 18, 42, 23, 4, 32, 14, 21, 3, 12, 19, 41

4. **Sandwich Prices** The prices (in dollars) of sandwiches at a restaurant are shown below. Use a frequency distribution and a histogram to organize the data. *(Spreadsheet at LarsonPrecalculus.com)*

   DATA 4.00, 4.00, 4.25, 4.50, 4.75, 4.25 5.95, 5.50, 5.50, 5.75, 6.25

B.2 **Finding Measures of Central Tendency** **In Exercises 5–8, find the mean, median, and mode of the data set.**

5. 6, 13, 8, 15, 9, 10, 8

6. 29, 36, 31, 38, 32, 31

7. 4, 11, 6, 13, 8, 6

8. 31, 38, 33, 40, 35

Choosing a Measure of Central Tendency **In Exercises 9 and 10, determine which measure of central tendency is the most representative of the data shown in the frequency distribution.**
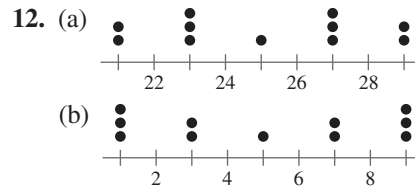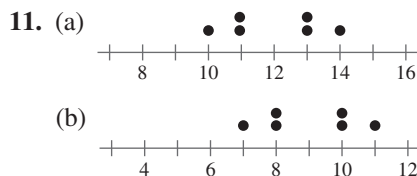
9.
Number	1	2	3	4	5	6	7	8	9
Frequency	3	2	1	0	0	0	1	2	3

10.
Number	1	2	3	4	5	6	7	8
Frequency	10	9	2	3	1	16	6	19

Finding Mean and Standard Deviation **In Exercises 11 and 12, line plots of sets of data are given. Find the mean and standard deviation of each set of data.**

11. (a)

(b)

12. (a)

(b)

Finding Mean, Variance, and Standard Deviation **In Exercises 13 and 14, find the mean ($\bar{x}$), variance (v), and standard deviation (σ) of the data set.**

13. 1, 2, 3, 4, 5, 6, 7   14. 1, 1, 1, 5, 5, 5

Using the Alternative Formula for Standard Deviation **In Exercises 15 and 16, use the alternative formula for standard deviation to find the standard deviation of the data set.**

15. 4, 5, 5, 6, 7, 7, 8   16. 3.3, 2.1, 4.5, 5.5, 6.0

Quartiles and Box-and-Whisker Plots **In Exercises 17 and 18, (a) find the lower and upper quartiles and (b) sketch a box-and-whisker plot for the data set.**

17. 23, 15, 14, 23, 13, 14, 13, 20, 12

18. 78.4, 76.3, 107.5, 78.5, 93.2, 90.3, 77.8, 37.1, 97.1, 75.5, 58.8, 65.6

Using a Normal Curve **In Exercises 19 and 20, determine the percent of the area under the normal curve represented by the shaded region.**

19.

20.
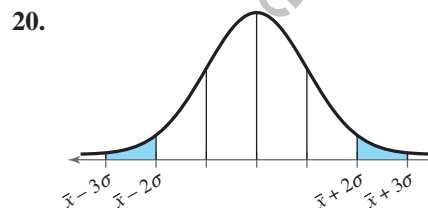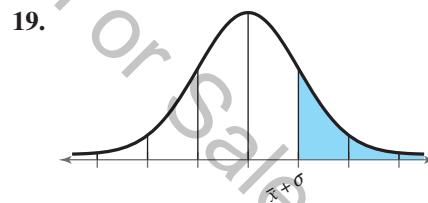
Finding a Normal Probability **In Exercises 21–24, a normal distribution has mean $\bar{x}$ and standard deviation σ. Find the indicated probability for a randomly selected x-value from the distribution.**

21. $P(x \leq \bar{x} + \sigma)$

22. $P(x \geq \bar{x} - \sigma)$

23. $P(21 \leq x \leq 41)$ when $\bar{x} = 33$ and $\sigma = 4$

24. $P(x \geq 25)$ when $\bar{x} = 33$ and $\sigma = 4$

Using the Standard Normal Table   **In Exercises 25 and 26, a normal distribution has a mean of 64 and a standard deviation of 7. Use the standard normal table to find the indicated probability for a randomly selected x-value from the distribution.**

25. $P(x \le 54)$ **26.** $P(x \ge 59)$

27. Fire Department   The time a fire department takes to arrive at the scene of an emergency is normally distributed with a mean of 6 minutes and a standard deviation of 1 minute.

(a) What is the probability that the fire department takes at most 8 minutes to arrive at the scene of an emergency?

(b) What is the probability that the fire department takes between 4 and 7 minutes to arrive at the scene of an emergency?

28. College Entrance Tests   Lisa and Ann took different college entrance tests. The scores on the test that Lisa took are normally distributed with a mean of 20 points and a standard deviation of 4.2 points. The scores on the test that Ann took are normally distributed with a mean of 500 points and a standard deviation of 90 points. Lisa scored 30 on her test, and Ann scored 610 on her test.

(a) Find the z-score for Lisa's test score.

(b) Find the z-score for Ann's test score.

(c) Which student scored better on her college entrance test? Explain your reasoning.

B.3 **Finding the Correlation Coefficient**   **In Exercises 29 and 30, you are given a set of data points and the least squares regression line for the data. Find the correlation coefficient r. How well does the model fit the data?**

29. $(-3, 2), (-2, 3), (-1, 4), (0, 6); y = 1.3x + 5.7$
30. $(-3, 4), (-1, 2), (1, 1), (3, 0); y = -0.65x + 1.75$

Finding the Sum of Square Differences   **In Exercises 31 and 32, you are given a set of data points and a linear model for the data. Find the sum of square differences for the linear model.**

31. $(0, 10), (1, 9), (2, 6), (3, 0)$
    $y = -3.3x + 11$
32. $(-1, -4), (1, -3), (2, 0), (4, 5), (6, 9)$
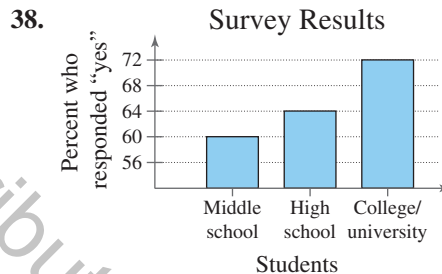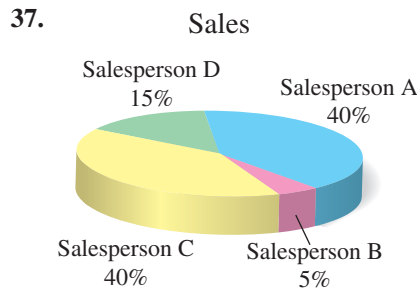 $y = 2.0x - 3$

Finding a Least Squares Regression Line   **In Exercises 33 and 34, find the least squares regression line for the points. Verify your answer with a graphing utility.**

33. $(0, 1), (1, 7), (2, 12)$
34. $(-1, 5), (0, 3), (1, 1), (2, -1)$

Finding a Least Squares Regression Parabola   **In Exercises 35 and 36, find the least squares regression parabola for the points. Verify your answer with a graphing utility.**

35. $(0, 1), (1, 8), (2, 1)$
36. $(-5, 4), (-4, 1), (-3, 0), (-2, 1)$

Misleading Graph   **In Exercises 37 and 38, explain why the graph could be misleading. Then redraw the graph so it is not misleading.**

37.

Sales

Salesperson D 15%
Salesperson A 40%
Salesperson C 40%
Salesperson B 5%

38.

Survey Results

Exploration

True or False?   **In Exercises 39–44, determine whether the statement is true or false. Justify your answer.**

39. The measure of central tendency that is most likely to be affected by the presence of an outlier is the mean.

40. It is possible for the mean, median, and mode of a data set to be equal.

41. Populations are collections of some of the outcomes, measurements, counts, or responses that are of interest.

42. Inferential statistics involves drawing conclusions about a sample using a population.

43. Data that are modeled by $y = 3.29x - 4.17$ have a negative linear correlation.

44. When the correlation coefficient for a linear regression model is close to -1, the regression line cannot be used to describe the data.

45. Reasoning   When the standard deviation of a data set of numbers is 0, what does this imply about the set?

46. Think About It   Construct a collection of numbers that has the following properties. If this is not possible, then explain why.

Mean $= 6$, median $= 6$, mode $= 4$

Chapter Test

Take this test as you would take a test in class. When you are finished, check your work against the answers given in the back of the book.

In Exercises 1–4, consider the following data set.

14, 16, 45, 7, 32, 32, 14, 16, 26, 46, 27, 43, 33, 27, 37, 12, 20, 40, 34, 16

1. Construct a line plot of the data.

2. Construct a stem-and-leaf plot of the data.

3. Construct a frequency distribution of the data.

4. Construct a histogram of the data.

In Exercises 5 and 6, find the mean, median, and mode of the data set.

5. 3, 10, 5, 12, 6, 7, 5

6. 32, 39, 34, 41, 35, 34

In Exercises 7 and 8, find the mean ($\bar{x}$), variance (v), and standard deviation (σ) of the data set.

7. 0, 1, 2, 3, 4, 5, 6

8. 2, 2, 2, 7, 7, 7

In Exercises 9 and 10, (a) find the lower and upper quartiles and (b) sketch a box-and-whisker plot for the data set.

9. 9, 5, 5, 5, 6, 5, 4, 12, 7, 10, 7, 11, 8, 9, 9

10. 25, 20, 22, 28, 24, 28, 25, 19, 27, 29, 28, 21

In Exercises 11 and 12, a normal distribution has a mean of 25 and a standard deviation of 3. Find the probability that a randomly selected x-value from the distribution is in the given interval.

11. Between 19 and 31

12. At most 22

In Exercises 13–15, consider the points

(0, 2),   (3, 4),   and   (4, 5).

13. Find the least squares regression line for the points. Verify your answer with a graphing utility.

14. Find the sum of square differences using the linear model you found in Exercise 13.

15. Find the correlation coefficient r. How well does the linear model fit the data?

16. Find the least squares regression parabola for the points $(-2, 9)$, $(-1, 4)$, and $(0, 2)$. Verify your answer with a graphing utility.

17. The guayule plant, which grows in the southwestern United States and in Mexico, is one of several plants used as a source of rubber. In a large group of guayule plants, the heights of the plants are normally distributed with a mean of 12 inches and a standard deviation of 2 inches.

    (a) What percent of the plants are taller than 17 inches?

    (b) What percent of the plants are at most 13 inches?

    (c) What percent of the plants are between 7 and 14 inches?

    (d) What percent of the plants are at least 3 inches taller than or at least 3 inches shorter than the mean height?

Proofs in Mathematics

Alternative Formula for Standard Deviation   *(p. B11)*

The standard deviation of $\{x_1, x_2, \ldots, x_n\}$ is

$$\sigma = \sqrt{\frac{x_1^2 + x_2^2 + \cdots + x_n^2}{n} - \bar{x}^2}.$$

Proof

To prove this formula, begin with the formula for standard deviation given in the definition on page B10.

$\sigma = \sqrt{v}$ — Square root of variance

$= \sqrt{\dfrac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n}}$ — Substitute definition of variance.

$= \sqrt{\dfrac{(x_1^2 - 2x_1\bar{x} + \bar{x}^2) + (x_2^2 - 2x_2\bar{x} + \bar{x}^2) + \cdots + (x_n^2 - 2x_n\bar{x} + \bar{x}^2)}{n}}$ — Square binomials.

$= \sqrt{\dfrac{(x_1^2 + x_2^2 + \cdots + x_n^2) - 2(x_1\bar{x} + x_2\bar{x} + \cdots + x_n\bar{x}) + \underbrace{(\bar{x}^2 + \bar{x}^2 + \cdots + \bar{x}^2)}_{n \text{ terms}}}{n}}$ — Regroup terms in numerator.

$= \sqrt{\dfrac{(x_1^2 + x_2^2 + \cdots + x_n^2)}{n} - \dfrac{2(x_1\bar{x} + x_2\bar{x} + \cdots + x_n\bar{x})}{n} + \dfrac{n\bar{x}^2}{n}}$ — Rewrite radicand as three fractions.

$= \sqrt{\dfrac{(x_1^2 + x_2^2 + \cdots + x_n^2)}{n} - \dfrac{2\bar{x}(x_1 + x_2 + \cdots + x_n)}{n} + \dfrac{n\bar{x}^2}{n}}$ — Factor $\bar{x}$ out of numerator of second fraction.

$= \sqrt{\dfrac{(x_1^2 + x_2^2 + \cdots + x_n^2)}{n} - 2\bar{x}^2 + \bar{x}^2}$ — Definition of $\bar{x}$; divide out n in third fraction.

$= \sqrt{\dfrac{(x_1^2 + x_2^2 + \cdots + x_n^2)}{n} - \bar{x}^2}$ — Simplify.

Not for Distribution or Sale

P.S. Problem Solving

1. U.S. History The years in which the first 20 states were admitted to the Union are as follows. Construct a stem-and-leaf plot of the data. *(Spreadsheet at LarsonPrecalculus.com)*

DATA 1788, 1787, 1788, 1816, 1792, 1812, 1788, 1788, 1817, 1788, 1787, 1788, 1789, 1803, 1787, 1790, 1788, 1796, 1791, 1788

2. Mayflower The known ages (in years) of adult male passengers on the *Mayflower* at the time of its departure were as follows. *(Spreadsheet at LarsonPrecalculus.com)*

DATA 21, 34, 29, 38, 30, 54, 39, 20, 35, 64, 37, 45, 21, 25, 55, 45, 40, 38, 38, 21, 21, 20, 34, 38, 50, 41, 48, 18, 32, 21, 32, 49, 30, 42, 30, 25, 38, 25, 20

(a) Construct a stem-and-leaf plot of the ages.

(b) Find the median age, and the range of the ages.

(c) According to one source, the age of passenger Thomas English was unknown at the time of the *Mayflower's* departure. What is the probability that he was 18–29 years old? Explain your reasoning.

(d) The first two rows of a *cumulative frequency distribution* for the data are shown below. The *cumulative frequency* for a given interval is the sum of the current frequency and all preceding frequencies. A histogram constructed from a cumulative frequency distribution is called a *cumulative frequency histogram*. Copy and complete the cumulative frequency distribution. Then construct a cumulative frequency histogram for the data.

Interval	Frequency	Cumulative Frequency
[15, 25)	9	9
[25, 35)	11	9 + 11 = 20
?	?	?

3. Organizing Data Construct a line plot, frequency distribution, and histogram for the data given in (a) Exercise 1 and (b) Exercise 2.

4. Constructing a Stem-and-Leaf Plot Construct a stem-and-leaf plot that has the same distribution of data as the histogram shown below. Explain your steps.
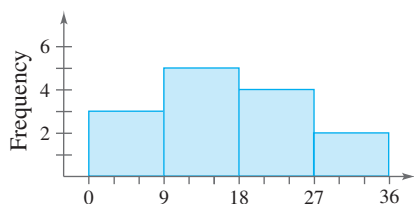
5. Think About It Two data sets have the same mean, interquartile range, and range. Is it possible for the box-and-whisker plots of such data sets to be different? Justify your answer by creating data sets that fit the situation.

6. Singers in a Chorus The box-and-whisker plots below show the heights (in inches) of singers in a chorus, according to their voice parts. A soprano part has the highest pitch, followed by alto, tenor, and bass, respectively. Draw a conclusion about voice parts and heights. Justify your conclusion.
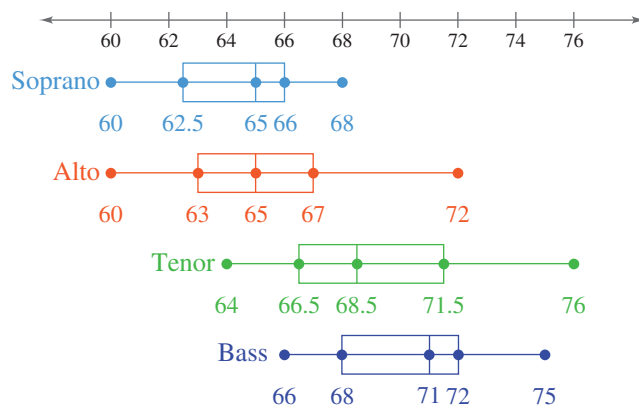
7. Equation of a Normal Curve A normal curve is defined by an equation of the form

$$y = \frac{1}{\sigma\sqrt{2\pi}}e^{-1/2[(x-\bar{x})/\sigma]^2}$$

where $\bar{x}$ is the mean and σ is the standard deviation of the corresponding normal distribution.

(a) Use a graphing utility to graph three equations of the given form, where $\bar{x}$ is held constant and σ varies.

(b) Describe the effect that the standard deviation has on the shape of a normal curve.

8. Height According to a survey by the National Center for Health Statistics, the heights of male adults in the United States are normally distributed with a mean of 69 inches and a standard deviation of 2.75 inches.

(a) What is the probability that 3 randomly selected U.S. male adults are all more than 6 feet tall?

(b) What is the probability that 5 randomly selected U.S. male adults are all between 65 and 75 inches tall?

9. Test Scores The scores of a mathematics exam given to 600 science and engineering students at a college had a mean and standard deviation of 235 and 28, respectively. Use Chebychev's Theorem to determine the intervals containing at least $\frac{3}{4}$ and at least $\frac{8}{9}$ of the scores. How would the intervals change when the standard deviation is 16?