Appendix B Concepts in Statistics

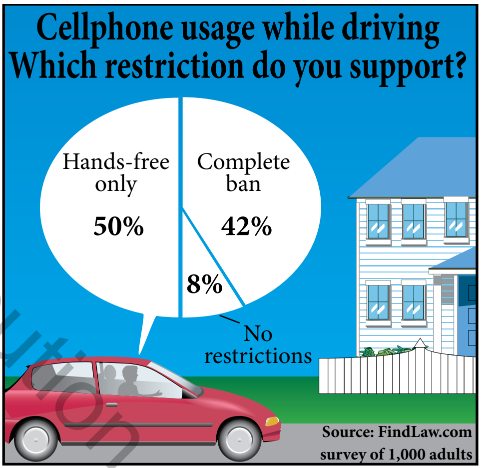
B.1 Representing Data

Line plots, stem-and-leaf plots, and histograms have many real-life applications. For example, in Exercise 17 on page B7, you will use a line plot to analyze gasoline prices.

- **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

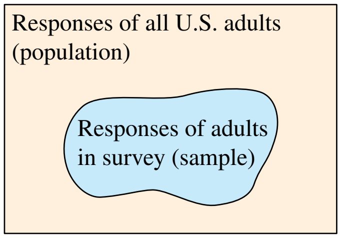
Statistics is the branch of mathematics that studies techniques for collecting, organizing, and interpreting information. **Data** consist of information that comes from observations, responses, counts, or measurements. Sometimes, the data are presented graphically, as illustrated by the circle graph below.



Source: FindLaw.com survey of 1,000 adults

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 example, in the survey represented above, the population is the collection of the responses of all U.S. adults, and the sample is the collection of the responses of the 1000 U.S. adults who participated in the survey. The Venn diagram below illustrates this.



HISTORICAL NOTE

The word *statistics* comes from the Latin *status*, which means "state." Using statistics dates back to ancient Babylonia, Egypt, and the Roman Empire. Census takers would collect data concerning the state, such as the numbers of occurrences of births and deaths.

In this example, the sample consists of about 420 responses that cellphone usage should be banned while driving, about 500 responses that only hands-free cellphone usage should be allowed, and about 80 responses that there should be no restrictions. The sample is a part, or subset, of the responses of all U.S. adults.

issarapong srirungpanich/Shutterstock.com

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 using a sample to draw conclusions about a population. Example 1 illustrates the difference between descriptive and inferential statistics.

EXAMPLE 1 Descriptive and Inferential Statistics

In a sample of 614 small business owners, 42% have a Facebook presence and find it valuable, and 30% have a Facebook presence but do not find it valuable. *(Source: Manta)*

a. Which part of the study represents descriptive statistics?

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

Solution

a. The statement "42% (of small business owners) have a Facebook presence and find it valuable, and 30% have a Facebook presence but do not find it valuable" represents descriptive statistics.

b. One possible conclusion you could draw from the study using inferential statistics is that more small business owners with a Facebook presence find it valuable than do not find it valuable.

Additional Example

In a sample of U.S. adults, 22% have donated money to an online crowdsourcing project.

a. Which part of the study represents descriptive statistics?

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

Answers

a. "22% (of U.S. adults) have donated money to an online crowdsourcing project."

b. *Sample answer:* Most U.S. adults have not donated to an online crowdsourcing project.

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

In a sample of college graduates, 86% said that college has been a good investment for them personally. *(Source: Pew Research Center)*

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** consist of numerical values. **Qualitative data** consist of attributes, labels, or nonnumerical values. For example, 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 sample data than a **census,** which is a survey of an entire population. Some sampling methods are listed below.

REMARK A stratified sample ensures that every segment of a population is represented.

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, such as selecting 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

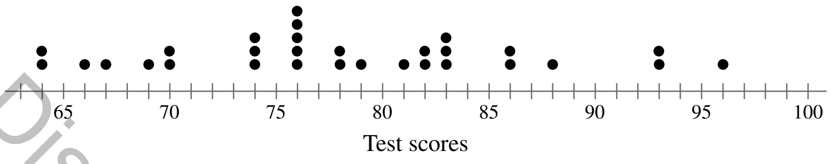
A **line plot** uses a portion of a real number line to order numbers. Line plots are especially useful for ordering small sets of numbers (about 50 or less) by hand.

EXAMPLE 2 Constructing a Line Plot

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

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 determining the least and greatest data values. For this data set, the least value 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 below. The line plot shows that 76 occurs with the greatest frequency.



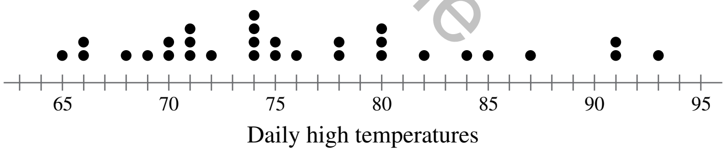
Checkpoint *Audio-video solution in English & Spanish at LarsonPrecalculus.com*

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

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 daily high temperatures (in degrees Fahrenheit) in a city during the month of June.



a. What is the range of daily high temperatures?

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

Solution

a. The line plot shows that the maximum daily high temperature was 93°F and the minimum daily 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 used 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 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 example, compare the test scores above with the following set of 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 from least to greatest.

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, construct a *double* stem-and-leaf plot as shown below. Note that there is one column of stems, with the leaves representing the units digits for the first set of test scores at the right of the stems and the leaves representing the units digits for the second set of test scores at the left of the stems.

Leaves (2nd Set)	*Stems*	*Leaves (1st Set)*
	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 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 set, 62 for 2nd set

The double stem-and-leaf plot shows that as a group, the test scores in the second set are higher than those in the first set.

DATA

Spreadsheet at LarsonPrecalculus.com

AK	9.4	MT	16.7
AL	15.3	NC	14.7
AR	15.7	ND	14.2
AZ	15.9	NE	14.4
CA	12.9	NH	15.9
CO	12.7	NJ	14.7
CT	15.5	NM	15.3
DE	16.4	NV	14.2
FL	19.1	NY	14.7
GA	12.4	OH	15.5
HI	16.1	OK	14.5
IA	15.8	OR	16.0
ID	14.3	PA	16.7
IL	13.9	RI	15.7
IN	14.3	SC	15.8
KS	14.3	SD	15.3
KY	14.8	TN	15.1
LA	13.6	TX	11.5
MA	15.1	UT	10.0
MD	13.8	VA	13.8
ME	18.3	VT	16.9
MI	15.4	WA	14.1
MN	14.3	WI	15.2
MO	15.4	WV	17.8
MS	14.3	WY	14.0

EXAMPLE 4 Using a Stem-and-Leaf Plot

The table at the left shows the percent of the population of each state that was at least 65 years old in 2014. Use a stem-and-leaf plot to organize the data. Then use the plot to identify the least and greatest data values. *(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 see that 9.4% is the least data value and 19.1% is the greatest data value.

Stems	*Leaves*
9	4
10	0
11	5
12	4 7 9
13	6 8 8 9
14	0 1 2 2 3 3 3 3 3 4 5 7 7 7 8
15	1 1 2 3 3 3 4 4 5 5 7 7 8 8 9 9
16	0 1 4 7 7 9
17	8
18	3
19	1

Key: 9|4 = 9.4

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

Use the stem-and-leaf plot in Example 4 to determine the number of states in which 16% or more of the population was at least 65 years old in 2014.

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

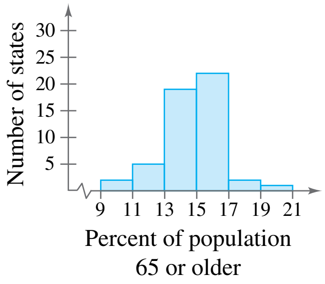
Histograms and Frequency Distributions

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

Interval	*Tally*
[9, 11)	||
[11, 13)	𝍸
[13, 15)	𝍸 𝍸 𝍸 ||||
[15, 17)	𝍸 𝍸 𝍸 𝍸 ||
[17, 19)	||
[19, 21)	|

Histogram



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.

EXAMPLE 5 Constructing a Histogram

See LarsonPrecalculus.com for an interactive version of this type of example.

A company has 48 sales representatives who sold the following numbers of units during the last quarter of 2016. 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. The least number is 100 and the greatest is 195, so you can use ten 10-unit intervals. Let the first interval be 100–109, the second interval be 110–119, and so on. Tally the data into the 10 intervals to obtain the frequency distribution shown at the left. Figure B.1 shows the histogram for the distribution.

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

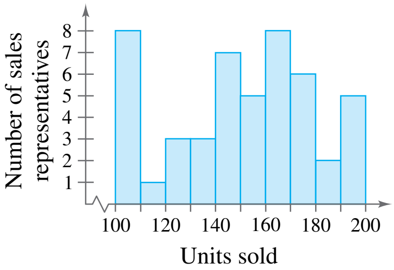


Figure B.1

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

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

Summarize (Section B.1)

1. Describe inferential and descriptive statistics *(page B2, Example 1)*.
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.

discpicture/Shutterstock.com

B.1 Exercises

See **CalcChat.com** for tutorial help and worked-out solutions to odd-numbered 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 using a sample to draw conclusions about a population.

5. ________ data consist of numerical values, whereas ________ data consist 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, researchers only select members of the population who are easily accessible.

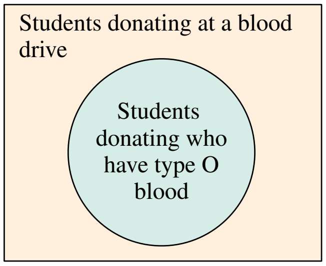
9. In a ________ sample, members of the population select themselves by volunteering.

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

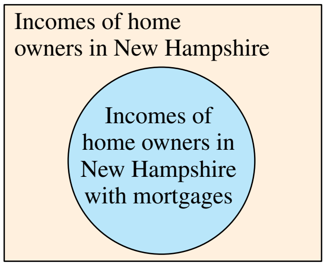
Skills and Applications

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

11.



12.



13. Financial Support In a survey of 750 parents, 31% plan to financially support their children through college graduation and 6% plan to financially support their children through the start of college. *(Source: Yahoo Financial)*

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

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

14. Superstition In a sample of 800 adults, 16% said that they are superstitious. *(Rasmussen Reports)*

(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 given 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)***

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

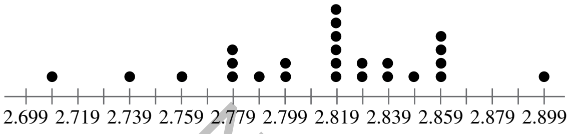
15. Construct a line plot for each quiz. For each quiz, which score occurred with the greatest frequency?

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 per gallon of unleaded regular gasoline from 25 different cities.

(a) In how many cities was the price of gasoline less than $2.80?

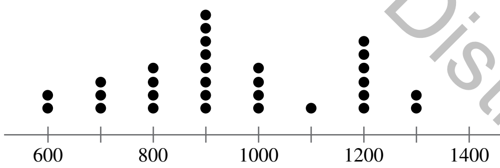
(b) What is the range of prices?



18. Cattle Weights The line plot shows the weights (to the nearest hundred pounds) of 30 cattle.

(a) How many cattle weighed about 800 to 1000 pounds?

(b) What is the range of weights?



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

DATA

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. Use the plot to identify the highest and lowest scores on the test.

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

21. Retirement Contributions The amounts (in dollars) that 35 employees contribute to their personal retirement plans are listed. Use a frequency distribution and a histogram to organize the data. *(Spreadsheet at LarsonPrecalculus.com)*

DATA

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 number of farms (in thousands) in each of the 50 states in 2015. Use a frequency distribution and a histogram to organize the data. *(Source: U.S. Dept. of Agriculture)*

DATA

Spreadsheet at LarsonPrecalculus.com

AK	1	AL	43	AR	44	AZ	20	CA	78
CO	34	CT	6	DE	3	FL	47	GA	41
HI	7	IA	88	ID	24	IL	74	IN	58
KS	60	KY	76	LA	27	MA	8	MD	12
ME	8	MI	52	MN	74	MO	97	MS	37
MT	28	NC	49	ND	30	NE	49	NH	4
NJ	9	NM	25	NV	4	NY	36	OH	74
OK	78	OR	35	PA	58	RI	1	SC	24
SD	31	TN	67	TX	242	UT	18	VA	45
VT	7	WA	36	WI	69	WV	21	WY	12

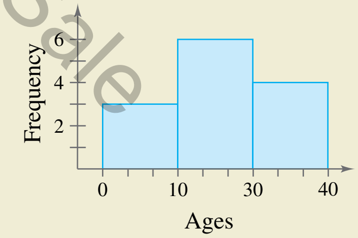
23. Meteorology The seasonal snowfall amounts (in inches) for Chicago, Illinois, for the years 1982 through 2014 are listed. (The amounts are listed in order by year.) How would you organize the data? Explain your reasoning. *(Source: National Weather Service)* *(Spreadsheet at LarsonPrecalculus.com)*

DATA

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, 30.1, 82.0, 50.7

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)	卌 |
[20, 30)	
[30, 40)	||||



Exploration

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. The ranges of data from two random samples from the same population must be equal.

27. Writing Describe how you could choose samples of students at your school using each of the five sampling methods listed on page B2.

issarapong srirungpanich/Shutterstock.com

B.2 Analyzing Data

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

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

Mean, Median, and Mode

It is often helpful to describe a data set by a single number that is most representative of the entire collection of numbers. Such a number is a **measure of central tendency.** The most commonly used measures are listed below.

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.** When no entry occurs more than once, the data set has no mode.

EXAMPLE 1 Finding Measures of Central Tendency

The annual incomes of 25 employees of a company are listed below. What are the mean, median, and mode of the incomes? *(Spreadsheet at LarsonPrecalculus.com)*

$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

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

To find the median, order the incomes.

$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

In this list, the median income is the middle income, $33,450. The income $12,500 occurs twice and 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 data set below.

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

In Example 1, note that the three measures of central tendency differ considerably. The mean is inflated by the two highest salaries and the mode is the least value. So, in this case, the median best represents a "typical" income.

REMARK The way in which the data is used can affect which measure best represents the data. For instance, in Example 1, the median best represents a "typical" income to a potential employee, but the mean is more useful to an accountant estimating a payroll budget.

Piccia Neri/Shutterstock.com

Choosing a Measure of Central Tendency

There are different ways to determine which measure of central tendency best represents a data set. In general, the measure that is close to most of the data is the most representative of the data set. For example, in both data sets below, the mean is 6, the median is 6, and the modes are 1 and 11.

Data set A: 1, 1, 1, 1, 1, 11, 11, 11, 11, 11

Data set B: 1, 1, 5.7, 5.8, 5.9, 6.1, 6.2, 6.3, 11, 11

The modes are the most representative measure of the data in set A, and the mean or median is the most representative measure of the data in set B.

EXAMPLE 2 Choosing Measures of Central Tendency

Determine 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	0	0	2	5	6	8	2	1	1

b.

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

c.

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

Solution

a. For these data, the mean is 5.4, the median is 5, and the mode is 6. Most of the data are close to the mean and the median, so the mean or median is 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). Most of the data are close to the modes, so the modes are the most representative measure.

c. For these data, the mean is approximately 3.45, the median is 3, and the mode is 3. Most of the data are close to the median and the mode, so the median or mode is the most representative measure.

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

Determine 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	4	20	3	0	0	0	4	20	3

b.

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

c.

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

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 the *variance* of the set and the *standard deviation* of the set.

Definitions of Variance and Standard Deviation

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

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

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

The greater the standard deviation of a data set, the more the numbers in the set *vary* from the mean. For example, each of the data sets below 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 Comparing Standard Deviations

The three data sets represented by the histograms in Figure B.2 all have a mean of 4. 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 mean of 4 and the numbers in data set *C* are the most dispersed from the mean. So, data set *A* has the least standard deviation and data set *C* has the greatest standard deviation.

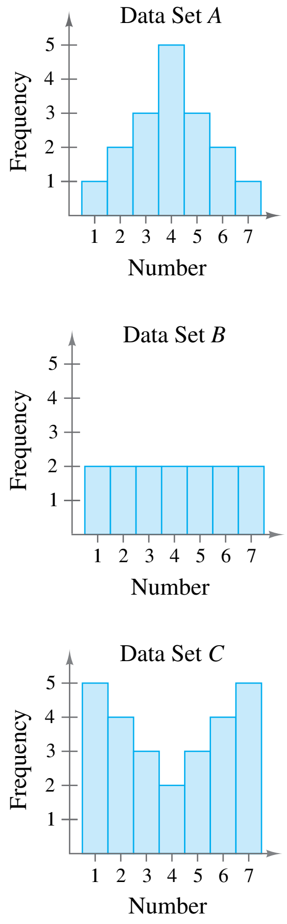
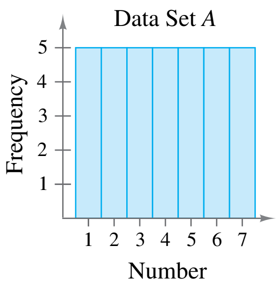
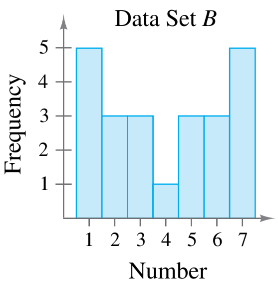


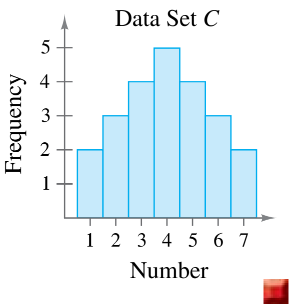
Figure B.2

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

The three data sets represented by the histograms shown below all have a mean of 4. Which data set has the least standard deviation? Which has the greatest?







EXAMPLE 4 Finding Standard Deviation

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

Solution Each data set 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 in the Checkpoint with Example 3.

The alternative formula below 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. Verify 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 set of numbers below.

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.

Additional Example

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

1, 3, 3, 5, 6, 6, 7, 7, 7, 9

Answer: About 2.29

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

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

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

A well-known theorem in statistics, called *Chebychev's Theorem*, states that the portion of any data set lying within k standard deviations of the mean is at least

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

where $k > 1$. So, at least 75% of the numbers in a data set must lie within two standard deviations of the mean, and at least $88.\overline{8}\%$ of the numbers must lie within three standard deviations of the mean. For most distributions, the percent of the numbers within k standard deviations of the mean is greater than the percent given by Chebychev's Theorem. For instance, in all three distributions shown in Example 3, 100% of the numbers lie within two standard deviations of the mean.

DATA

Spreadsheet at LarsonPrecalculus.com

AK	9.75	MT	8.05
AL	7.25	NC	7.25
AR	8.00	ND	7.25
AZ	8.05	NE	9.00
CA	10.00	NH	7.25
CO	8.31	NJ	8.38
CT	9.60	NM	7.50
DC	10.50	NV	8.25
DE	8.25	NY	9.00
FL	8.05	OH	8.10
GA	7.25	OK	7.25
HI	8.50	OR	9.25
IA	7.25	PA	7.25
ID	7.25	RI	9.60
IL	8.25	SC	7.25
IN	7.25	SD	8.55
KS	7.25	TN	7.25
KY	7.25	TX	7.25
LA	7.25	UT	7.25
MA	10.00	VA	7.25
MD	8.25	VT	9.60
ME	7.50	WA	9.47
MI	8.50	WI	7.25
MN	9.00	WV	8.75
MO	7.65	WY	7.25
MS	7.25		

EXAMPLE 6 Describing a Distribution

The table at the left shows the minimum wages of the 50 states and the District of Columbia in 2016. Find the mean and standard deviation of the data. What percent of the data values lie within one standard deviation of the mean? *(Source: U.S. Department of Labor)*

Solution Begin by entering the numbers into a graphing utility. Then use the *one-variable statistics* feature to obtain $\bar{x} \approx 8.12$ and $\sigma \approx 0.94$. The interval that contains all numbers that lie within one standard deviation of the mean is

$$[8.12 - 0.94, 8.12 + 0.94] \quad \text{or} \quad [7.18, 9.06].$$

From the table, you can see that all but nine of the data values (about 82%) lie in this interval.

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

In Example 6, what percent of the data values lie within two standard deviations of the mean?

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

Additional Example

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

26, 28, 11, 27, 15, 17, 8, 32, 36, 38, 4, 16

Answer: Lower quartile: 13,
upper quartile: 30

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

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

Solution Begin by ordering the data.

$$\underbrace{12, 13, 14,}_{\text{1st } 25\%} \quad \underbrace{16, 16, 18,}_{\text{2nd } 25\%} \quad \underbrace{20, 24, 24,}_{\text{3rd } 25\%} \quad \underbrace{26, 27, 42}_{\text{4th } 25\%}$$

The median of the entire data set is $(18 + 20)/2 = 19$. The median of the six numbers that are less than 19 is $(14 + 16)/2 = 15$. So, the lower quartile is 15. The median of the six numbers that are greater than 19 is 25. So, the upper quartile is $(24 + 26)/2 = 25$.

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

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

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. The interquartile range of the data set in Example 7 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 in the figure 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 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.

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, use it to recreate the box-and-whisker plot shown above for the data set in Example 7.

EXAMPLE 8 Sketching Box-and-Whisker Plots

See LarsonPrecalculus.com for an interactive version of this type of example.

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

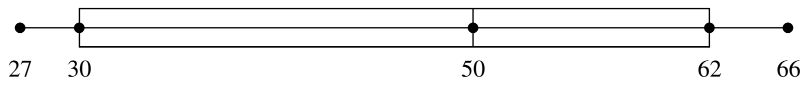
a. 27, 28, 30, 42, 45, 50, 50, 61, 62, 64, 66

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

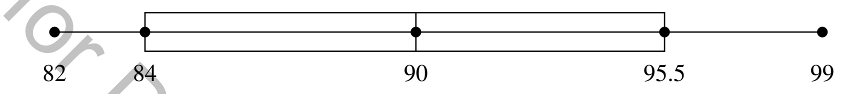
c. 11, 13, 13, 15, 17, 18, 20, 24, 24, 27

Solution

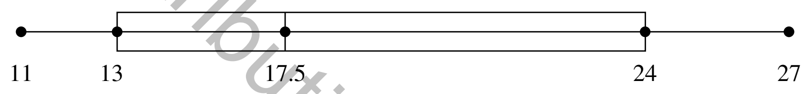
a. The median is 50. The lower quartile is 30 (the median of the first five numbers). The upper quartile is 62 (the median of the last five numbers). See the plot below.



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 the plot below.



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 the plot below.



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

Sketch a box-and-whisker plot for the data set: 2, 7, 9, 38, 44, 54, 56, 62, 79, 93.

Normal Distributions

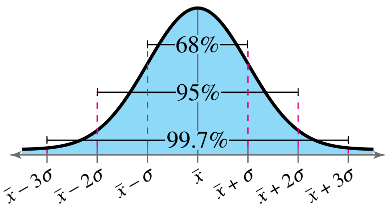
Recall that a **normal distribution** is modeled by a bell-shaped curve. This is called a **normal curve** and it is symmetric with respect to the mean.

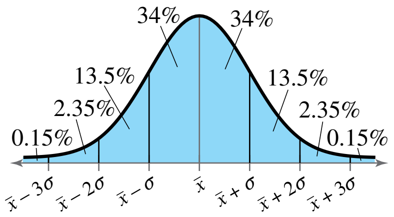
REMARK In Section 3.5, you studied Gaussian models, which are examples of normal distributions.

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.





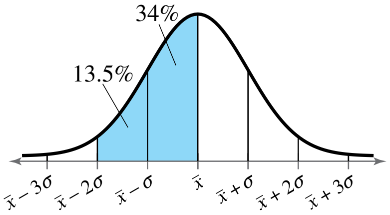
EXAMPLE 9 Finding a Normal Probability

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 at the right.



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

✓ ***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 levels 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 levels between 158 and 186 milligrams per deciliter?

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

Solution

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

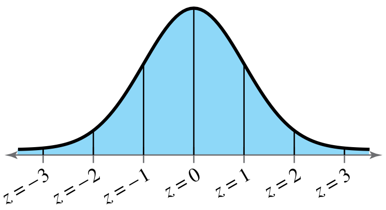
b. A level of 158 milligrams per deciliter is one standard deviation to the left of the mean. So, the percent of the levels that are desirable is about 0.15% + 2.35% + 13.5% = 16%.

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

In Example 10, about what percent of the women have levels 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.

REMARK In the table, the value .0000+ means "slightly more than 0" and the value 1.0000− means "slightly less than 1."

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

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

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.

A seal sanctuary has protected beaches where harbor seals give birth to their young in their natural habitat.

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 recorded 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 $P(x \le 50) \approx P(z \le -1.6) = 0.0548$. So, the probability that the scientists observed at most 50 seals during a survey is about 0.0548.

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 data set *(page B8)*. For an example of finding measures of central tendency, see Example 1.
2. Explain how to determine the measure of central tendency that best represents a data set *(page B9)*. For an example of determining the measures of central tendency that best represent data sets, see Example 2.
3. Explain how to find the standard deviation of a data set *(pages B10 and B11)*. For examples of finding the standard deviations of data sets, 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 *(page B14)*. For examples of using normally distributed data, see Examples 9–11.

Jody/Shutterstock.com

B.2 Exercises

See **CalcChat.com** for tutorial help and worked-out solutions to odd-numbered exercises.

Vocabulary: Fill in the blanks.

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

2. When there is an even number of data values in a data set, 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 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. Quartiles are represented graphically by 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.

Skills and Applications

Finding Measures of Central Tendency

In Exercises 11–16, find the mean, median, and mode(s) of the data set.

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

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

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

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

15. 5, 6, 12, 7, 14, 9, 7, 7, 6

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

17. Electric Bills A homeowner's monthly electric bills for a year are listed. What are the mean and median of the data? *(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 records the number of cars rented each day for a week. The results are listed. 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
Sunday	180		

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

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

Spreadsheet at LarsonPrecalculus.com

20. Sports The table shows the numbers of games in which a baseball player had 0, 1, 2, 3, and 4 hits during the last 50 games.

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

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

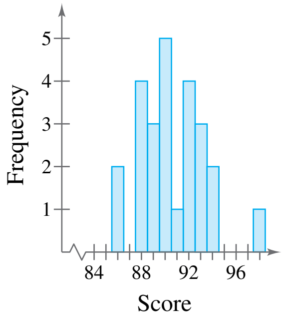
(b) The player had a total of 200 at-bats in the 50 games. Determine the player's batting average.

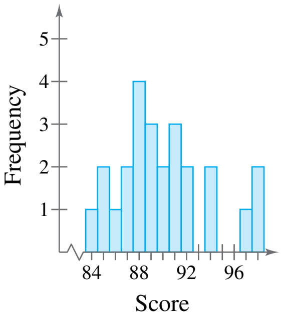
21. Test Scores A professor records the students' scores for a 100-point exam. The results are listed. *(Spreadsheet at LarsonPrecalculus.com)*

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. Test Scores The histograms represent the test scores of two classes of a college course in mathematics. Which histogram shows the lesser standard deviation? Explain.

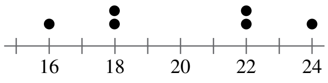




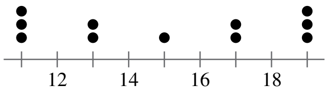
Finding Mean and Standard Deviation **In Exercises 23 and 24, each line plot represents a data set. Find the mean and standard deviation of each data set.**

23. (a)

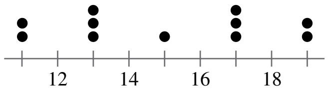
(b)



24. (a)



(b)



Finding Mean, Variance, and Standard Deviation **In Exercises 25–30, find the mean ($\bar{x}$), variance (v), and standard deviation (σ) of the data set.**

25. 4, 10, 8, 2

26. 2, 12, 4, 7, 5

27. 0, 1, 1, 2, 2, 2, 3, 3, 4

28. 2, 2, 2, 2, 2, 2

29. 42, 50, 61, 47, 56, 68

30. 1.2, 0.6, 2.8, 1.7, 0.9

Using the Alternative Formula for Standard Deviation **In Exercises 31–36, use the alternative formula for standard deviation to find the standard deviation of the data set.**

31. 3, 5, 7, 7, 18, 8

32. 24, 20, 38, 15, 52, 33, 46

33. 246, 336, 473, 167, 219

34. 5.1, 6.7, 4.5, 10.2, 9.1

35. 8.6, 6.4, 2.9, 5.0, 6.7

36. 9.2, 10.6, 7.2, 4.3, 7.0

Quartiles and Box-and-Whisker Plots **In Exercises 37–40, find the lower and upper quartiles and sketch a box-and-whisker plot for each data set.**

37. 11, 10, 11, 14, 17, 16, 14, 11, 8, 14, 20

38. 46, 48, 48, 50, 52, 47, 51, 47, 49, 53

39. 19, 12, 14, 9, 14, 15, 17, 13, 19, 11, 10, 19

40. 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

41. Price of Gold The data represent the average prices of gold (in dollars per troy ounce) for the years 1996 through 2015. Use a graphing utility to find the mean, variance, and standard deviation of the data. What percent of the data lies within one standard deviation of the mean? *(Source: U.S. Bureau of Mines and U.S. Geological Survey)* *(Spreadsheet at LarsonPrecalculus.com)*

389, 332, 295, 280, 280,
272, 311, 365, 411, 446,
606, 699, 874, 975, 1227,
1572, 1673, 1415, 1269, 1170

42. Product Lifetime A manufacturer redesigns a machine part to increase its average lifetime. The two data sets list the lifetimes (in months) of 20 randomly selected parts of each design. *(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

(a) Construct a box-and-whisker plot for each data set.

(b) Explain the differences between the box-and-whisker plots you constructed in part (a).

Piccia Neri/Shutterstock.com

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 25

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. $P(x \leq 71)$

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 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. Some quantitative data sets do not have medians.

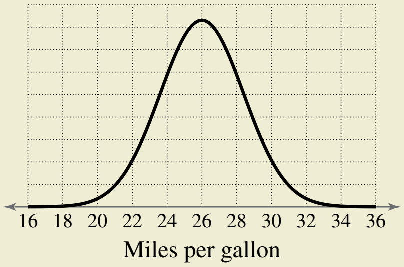
60. About one quarter of the data in a set is less than the lower quartile.

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 represents the ***n*th 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 model 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.



(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.

66. **Reasoning**

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

(b) Make a conjecture about how the measures of central tendency of a data set change when you add a constant k to each data value.

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.

Mean $= 6$
Median $= 4$
Mode $= 4$

B.3 Modeling Data

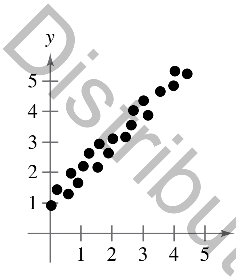
The method of least squares provides a way of creating mathematical models for data sets. For example, in Exercise 21 on page B25, you will find the least squares regression line for the average annual costs of healthcare for a family of four in the United States from 2012 through 2015.

- **Use the correlation coefficient to measure how well a model fits a data set.**
- **Use the sum of the squared differences to measure how well a model fits a data set.**
- **Find the least squares regression line for a data set.**
- **Find the least squares regression parabola for a data set.**
- **Analyze misleading graphs.**

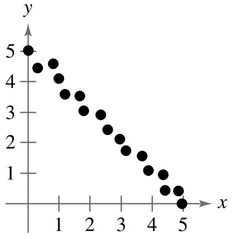
Throughout this text, you have been using or finding models for data sets. For example, in Section 1.10, you determined how closely a given model represents a data set, and you used the *regression* feature of a graphing utility to find a linear model for a data set. This section expands upon the concepts learned in Section 1.10 regarding the use of least squares regression to fit models to data.

Correlation

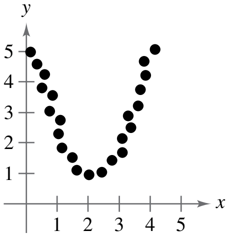
A **correlation** is a relationship between two variables. A scatter plot is helpful in determining whether a correlation exists between two variables. The scatter plots below illustrate several types of correlations.



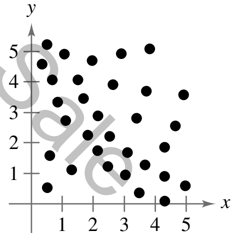
Positive linear correlation: y tends to increase as x increases.



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



Nonlinear correlation



No correlation

REMARK Recall from Section 1.10 that the correlation coefficient can be used to determine whether a linear model is a good fit for a data set.

The **correlation coefficient** r gives a measure of the strength and direction of a linear correlation between the two variables. The range of the possible values 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. To review summation notation, see Section 9.1.

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

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

Christian Delbert/Shutterstock.com

Sum of the Squared Differences

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

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

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

EXAMPLE 1 Finding the Sum of the Squared Differences

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

DATA

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

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

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

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

REMARK In Example 1, note that the sum of the squared differences is relatively large, so the model is likely not a good fit for the data set.

✓ ***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 the squared 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 the 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, solve the system below for a and b.

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

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

EXAMPLE 2 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 a table to find the coefficients and constants of the system of equations for the least squares regression line.

x	y	xy	x^2
-3	0	0	9
-1	1	-1	1
0	2	0	0
2	3	6	4
$\sum_{i=1}^{n} x_i = -2$	$\sum_{i=1}^{n} y_i = 6$	$\sum_{i=1}^{n} x_i y_i = 5$	$\sum_{i=1}^{n} x_i^2 = 14$

Apply the system for the least squares regression line with $n = 4$.

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

Solving this system of equations produces

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

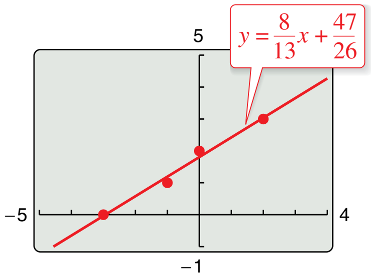


Figure B.3

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 $(2, 4)$.

Least squares regression analysis has a wide variety of real-life applications. For example, in Exercise 23, you will find the least squares regression line that models the demand for a tool at a hardware retailer as a function of price.

EXAMPLE 3 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 $n = 4$,

$$\begin{aligned} r &= \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{\sqrt{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}\sqrt{n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2}} \\ &= \frac{4(5) - (-2)(6)}{\sqrt{4(14) - (-2)^2}\sqrt{4(14) - 6^2}} \\ &\approx 0.992. \end{aligned}$$

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?

Golubovy/Shutterstock.com

Least Squares Regression Parabolas

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 system below for a, b, and c.

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

REMARK You may recall finding least square regression lines and parabolas in Section 7.2 (Exercises 54–58) and Section 7.3 (Exercises 67–70), respectively. In these exercises, the systems of equations were given because you had not yet studied summation notation.

EXAMPLE 4 Finding a Least Squares Regression Parabola

See LarsonPrecalculus.com for an interactive version of this type of example.

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

Solution Begin by constructing a table, as shown below.

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

Additional Example

Find the least squares regression parabola for the points (1, 2), (2, 4), (3, 8), and (4, 15).

Answer: $y = \frac{5}{4}x^2 - \frac{39}{20}x + \frac{11}{4}$

Use the table to find the sums needed to write the system of equations.

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

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

Apply the system for the least squares regression parabola with $n = 4$.

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

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.

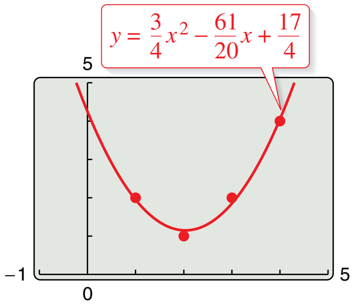


Figure B.4

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

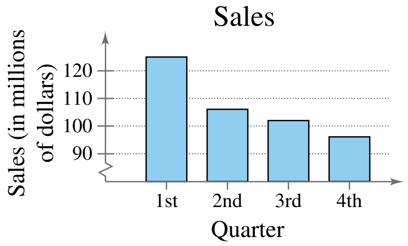
Find the least squares regression parabola for the points $(-3, -2)$, $(-2, 0)$, $(-1, 1)$, and $(1, 0)$.

Analyze Misleading Graphs

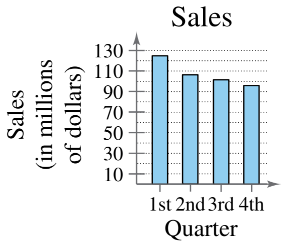
A *misleading graph* is a graph that is not drawn appropriately. This type of graph can misrepresent data and lead to false conclusions. For example, the bar graph below has a break in the vertical axis, which gives the impression that first quarter sales were disproportionately more than sales in other quarters.

REMARK To recognize when a graph is trying to deceive or mislead, ask yourself these questions:

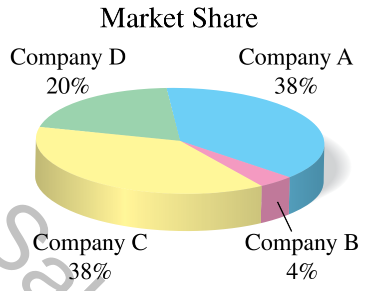
- Does the graph have a title?
- Does the graph need a key?
- Are the numbers of the scale evenly spaced?
- Are all the axes or sections of the graph labeled?
- Does the scale begin at zero? If not, is there a break?
- Are all the components of the graph, such as the bars, the same size?



One way to redraw this graph so it is not misleading is to show it without the break, as shown below. Notice that in the redrawn graph, the differences in the quarterly sales do not appear to be as dramatic.



Another example of a misleading graph is the circle graph at the right. This graph is shown at an angle that makes the market share of Company C appear larger than it is. Redrawing the graph as shown in Figure B.5 makes it clearer that Company A and Company C have equal market shares.



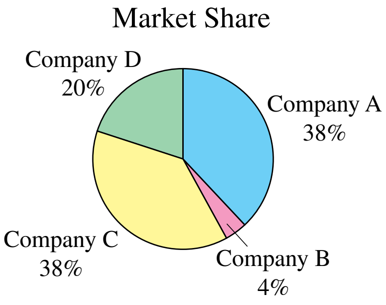


Figure B.5

Summarize (Section B.3)

1. Explain how to use the correlation coefficient to measure how well a model fits a data set *(page B20)*. For an example of finding a correlation coefficient, see Example 3.
2. Explain how to use the sum of the squared differences to measure how well a model fits a data set *(page B21)*. For an example of finding the sum of the squared differences for a linear model, see Example 1.
3. List the steps involved in finding the least squares regression line for a data set *(page B21)*. For an example of finding the least squares regression line for a data set, see Example 2.
4. List the steps involved in finding the least squares regression parabola for a data set *(page B23)*. For an example of finding the least squares regression parabola for a data set, see Example 4.
5. Give examples of misleading graphs *(page B24)*.

B.3 Exercises

See **CalcChat.com** for tutorial help and worked-out solutions to odd-numbered exercises.

Vocabulary: Fill in the blanks.

1. A ________ is a relationship between two variables.

2. A graphing utility uses the ________ of ________ ________ to find a mathematical model for a data set.

3. The sum of the ________ ________ measures how well a model fits a set of data points.

4. The ________ ________ ________ line for a data set is the linear model that has the least sum of the squared differences.

Skills and Applications

Finding the Sum of the Squared Differences **In Exercises 5–12, find the sum of the squared differences for the given data points and the model.**

5. (0, 2), (1, 1), (2, 2), (3, 4), (5, 6), $y^* = 0.8x + 2$

6. (0, 7), (2, 5), (3, 3), (4, 3), (6, 0), $y^* = -1.2x + 7$

7. (−3, −1), (−1, 0), (0, 2), (2, 3), (4, 4)
$y^* = 0.5x + 0.5$

8. (−2, 6), (−1, 4), (0, 2), (1, 1), (2, 1)
$y^* = -1.7x + 2.7$

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. Use a graphing utility to verify your answer.**

13. (−4, 1), (−3, 3), (−2, 4), (−1, 6)

14. (0, −1), (2, 0), (4, 3), (6, 5)

15. (−3, 18), (−1, 9), (2, 3), (4, −8)

16. (0, −1), (2, 1), (3, 2), (5, 3)

17. Finding a Correlation Coefficient Find the correlation coefficient *r* for the data given in Exercise 13. How well does the linear model fit the data?

18. Finding a Correlation Coefficient Find the correlation coefficient *r* for the data given in Exercise 14. How well does the linear model fit the data?

19. Finding a Correlation Coefficient Find the correlation coefficient *r* for the data given in Exercise 15. How well does the linear model fit the data?

20. Finding a Correlation Coefficient Find the correlation coefficient *r* for the data given in Exercise 16. How well does the linear model fit the data?

21. Healthcare Costs

The average annual costs of healthcare for a family of four in the United States from 2012 through 2015 are represented by the ordered pairs (x, y), where x represents the year, with $x = 12$ corresponding to 2012, and y represents the average cost (in thousands of dollars). Find the least squares regression line for the data. What is the sum of the squared differences? *(Source: Milliman)*

(12, 20.73), (13, 22.03), (14, 23.22), (15, 24.67)

22. 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 the squared differences? *(Source: U.S. National Center for Education Statistics)*

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

23. 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?

Christian Delbert/Shutterstock.com

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

(c) How much fertilizer should the agronomist apply to achieve a yield of 70 bushels per acre?

Finding a Least Squares Regression Parabola In Exercises 25–32, find the least squares regression parabola for the points. Use a graphing utility to verify your answer.

25. (0, 0), (2, 4), (4, 2)

26. $(-2, 6)$, $(-1, 2)$, (1, 3)

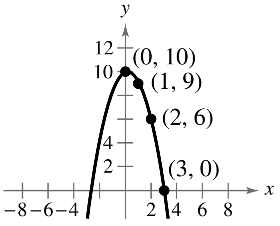
27. $(-1, 4)$, (0, 2), (1, 0), (3, 4)

28. $(-3, -1)$, $(-1, 2)$, (1, 2), (3, 0)

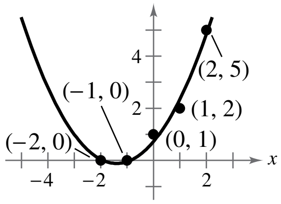
29. $(-2, 18)$, $(-1, 9)$, (0, 4), (1, 3), (2, 6)

30. $(-2, -17)$, (0, 0), (2, 4), $(4, -3)$, $(6, -25)$

31.

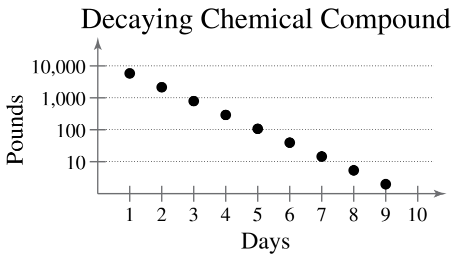


32.

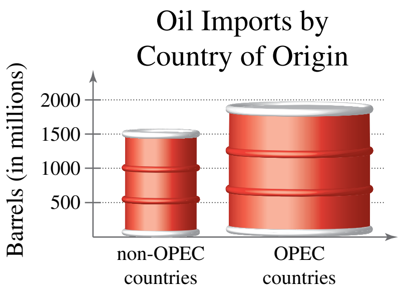


Misleading Graph In Exercises 33 and 34, explain why the graph could be misleading.

33.



34.



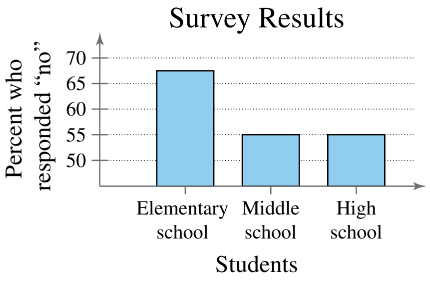
Exploration

True or False? In Exercises 35 and 36, determine whether the statement is true or false. Justify your answer.

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

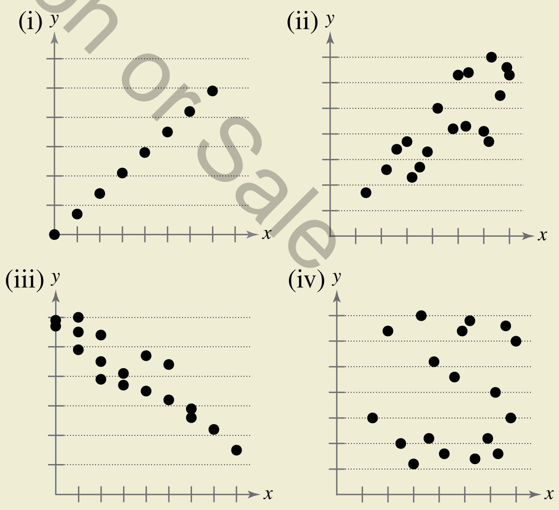
36. A linear regression model with a positive correlation coefficient has a graph with a slope that is greater than 0.

37. Error Analysis Describe the error in interpreting the graph.



The percent of elementary school students who responded "no" is more than twice the percent of middle school students who responded "no." ✗

38. HOW DO YOU SEE IT? Match each correlation coefficient r with the corresponding scatter plot. Explain your reasoning. [The scatter plots are labeled (i), (ii), (iii), and (iv).]

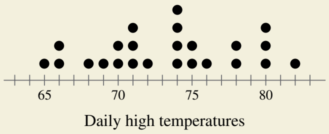


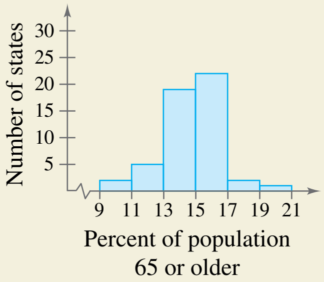
(a) $r = 0.04$ (b) $r = 1$

(c) $r = -0.92$ (d) $r = 0.81$

39. 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.

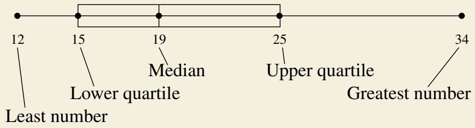
Chapter Summary

<table>
<tr><th></th><th>What Did You Learn?</th><th>Explanation/Examples</th><th>Review Exercises</th></tr>
<tr><td>Section B.1</td>
<td>Understand terminology associated with statistics <i>(p. B1)</i>, use line plots to order and analyze data <i>(p. B3)</i>, use stem-and-leaf plots to organize and compare data <i>(p. B4)</i>, and use histograms to represent frequency distributions <i>(p. B5)</i>.</td>
<td>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.
Line Plot


Daily high temperatures
Stem-and-Leaf Plot
<table>
<tr><th><i>Stems</i></th><th><i>Leaves</i></th></tr>
<tr><td>5</td><td>8</td></tr>
<tr><td>6</td><td>4 4 6 9</td></tr>
<tr><td>7</td><td>0 0 4 4 4 6 6 6 6 6 8 8 9</td></tr>
<tr><td>8</td><td>1 2 2 3 3 3 6 6 8</td></tr>
<tr><td>9</td><td>3 3 6</td></tr>
</table>
Key: 5|8 = 58
Frequency Distribution
<table>
<tr><th><i>Interval</i></th><th><i>Tally</i></th></tr>
<tr><td>[9, 11)</td><td>||</td></tr>
<tr><td>[11, 13)</td><td>𝍸</td></tr>
<tr><td>[13, 15)</td><td>𝍸 𝍸 𝍸 ||||</td></tr>
<tr><td>[15, 17)</td><td>𝍸 𝍸 𝍸 𝍸 ||</td></tr>
<tr><td>[17, 19)</td><td>||</td></tr>
<tr><td>[19, 21)</td><td>|</td></tr>
</table>
Histogram

</td>
<td>1–4</td></tr>
<tr><td rowspan="3">Section B.2</td>
<td>Find and interpret the mean, median, and mode of a data set <i>(p. B8),</i> and determine the measure of central tendency that best represents a data set <i>(p. B9).</i></td>
<td>The mean of n numbers is the sum of the numbers divided by n.
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.
The mode of n numbers is the number that occurs most frequently.</td>
<td>5–10</td></tr>
<tr><td>Find the standard deviation of a data set <i>(p. B10).</i></td>
<td>The variance of a set of numbers $\{x_1, x_2, . . . , x_n\}$ with a mean of $\bar{x}$ is

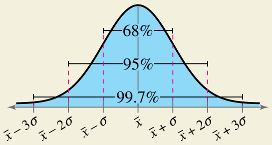
$$v = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdot \cdot \cdot + (x_n - \bar{x})^2}{n}$$

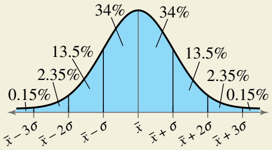
and the standard deviation is $\sigma = \sqrt{v}$.
Alternative Formula for Standard Deviation

$$\sigma = \sqrt{\frac{x_1^2 + x_2^2 + \cdot \cdot \cdot + x_n^2}{n} - \bar{x}^2}$$
</td>
<td>11–16</td></tr>
<tr><td>Create and use box-and-whisker plots <i>(p. B13).</i></td>
<td>Box-and-Whisker Plot

</td>
<td>17, 18</td></tr>
</table>

<table>
<tr><th></th><th>What Did You Learn?</th><th>Explanation/Examples</th><th>Review Exercises</th></tr>
<tr><td>Section B.2</td><td>Interpret normally distributed data *(p. B14)*.</td><td>

Normal distribution with mean $\bar{x}$ and standard deviation σ:





Using the formula $z = (x - \bar{x})/\sigma$ transforms an x-value from a normal distribution with mean $\bar{x}$ and standard deviation σ into a corresponding z-score having a standard normal distribution.

</td><td>19–30</td></tr>
<tr><td rowspan="4">Section B.3</td><td>Use the correlation coefficient to measure how well a model fits a data set *(p. B20)*.</td><td>

The correlation coefficient r gives a measure of how well a model fits a data set. A formula for r is

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

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

</td><td>31, 32</td></tr>
<tr><td>Use the sum of the squared differences to measure how well a model fits a data set *(p. B21)*.</td><td>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 the squared differences.</td><td>33, 34</td></tr>
<tr><td>Find the least squares regression line *(p. B21)* and the least squares regression parabola *(p. B23)* for data sets.</td><td>

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, solve the system below for a and b.

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

To find the least squares regression parabola $y = ax^2 + bx + c$ for the points $(x_1, y_1), (x_2, y_2), . . . , (x_n, y_n)$ algebraically, solve the system below for a, b, and c.

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

</td><td>35–38</td></tr>
<tr><td>Analyze misleading graphs *(p. B24)*.</td><td>A misleading graph is a graph that is not drawn appropriately. This type of graph can misrepresent data and lead to false conclusions. (See page B24 for examples.)</td><td>39, 40</td></tr>
</table>

Review Exercises

See **CalcChat.com** for tutorial help and worked-out solutions to odd-numbered exercises.

B.1 Constructing a Line Plot **In Exercises 1 and 2, construct a line plot for 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 for the data. *(Source: Major League Baseball) (Spreadsheet at LarsonPrecalculus.com)*

DATA 18, 4, 42, 33, 7, 19, 20, 9, 29, 15, 17, 15, 16, 2, 23, 6, 9, 44, 6, 7

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(s) 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, 33, 31

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.

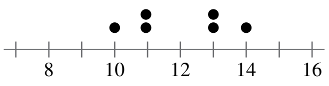
Number	1	2	3	4	5	6	7	8	9
Frequency	3	2	5	6	2	1	0	0	1

10.

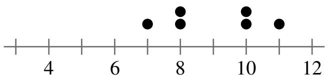
Number	1	2	3	4	5	6	7	8
Frequency	0	5	12	1	1	12	3	1

Finding Mean and Standard Deviation **In Exercises 11 and 12, each line plot represents a data set. Find the mean and standard deviation of each data set.**

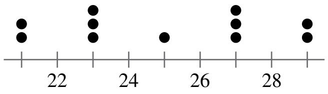
11. (a)



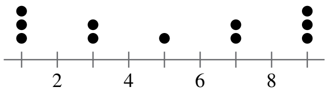
(b)



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. 2, 5, 5, 7, 7, 9, 11

16. 9.3, 8.5, 12.0, 11.6, 10.1

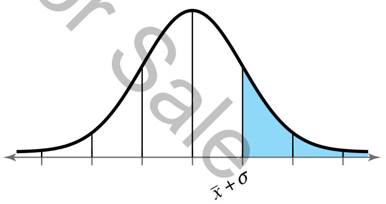
Quartiles and Box-and-Whisker Plots **In Exercises 17 and 18, find the lower and upper quartiles and sketch a box-and-whisker plot for the data set.**

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

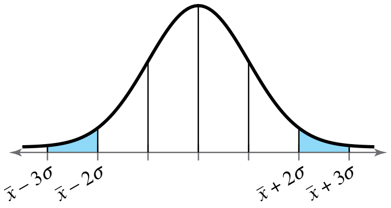
18. 78.4, 76.3, 107.5, 78.5, 93.2, 90.3, 77.8, 37.1, 97.1, 75.5, 58.8, 65.6

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 25–28, 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. $P(x \leq 75)$ **26.** $P(x \geq 66)$

27. $P(52 \leq x \leq 59)$

28. $P(x \leq 58 \text{ or } x \geq 70)$

29. **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?

30. **College Entrance Tests** Two students took different college entrance tests. The scores on the test that the first student took are normally distributed with a mean of 20 points and a standard deviation of 4.2 points. The scores on the test that the second student took are normally distributed with a mean of 500 points and a standard deviation of 90 points. The students scored 30 and 610 on their tests, respectively.

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

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

(c) Which student scored better on their college entrance test? Explain.

B.3 **Finding the Correlation Coefficient** **In Exercises 31 and 32, find the correlation coefficient *r* of the data set and describe the correlation.**

31. $(-3, 2), (-2, 3), (-1, 4), (0, 6)$

32. $(-3, 4), (-1, 2), (1, 1), (3, 0)$

Finding the Sum of the Squared Differences **In Exercises 33 and 34, find the sum of the squared differences for the set of data points and the model.**

33. $(0, 11), (1, 9), (2, 5), (3, 0)$

$y^* = -3.7x + 12$

34. $(-1, -4), (1, -3), (2, 0), (4, 5), (6, 9)$

$y^* = 2.0x - 3$

Finding a Least Squares Regression Line **In Exercises 35 and 36, find the least squares regression line for the points. Use a graphing utility to verify your answer.**

35. $(0, 1), (1, 7), (2, 12)$

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

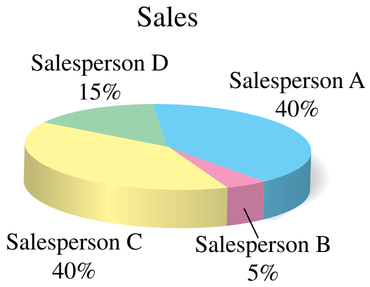
Finding a Least Squares Regression Parabola **In Exercises 37 and 38, find the least squares regression parabola for the points. Use a graphing utility to verify your answer.**

37. $(0, 1), (1, 8), (2, 1)$

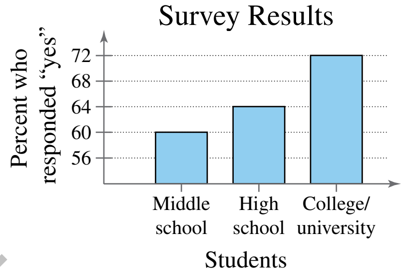
38. $(-5, 5), (-4, 1), (-3, 0), (-2, 1), (-1, 4)$

Misleading Graph **In Exercises 39 and 40, explain why the graph could be misleading.**

39.



40.



Exploration

True or False? **In Exercises 41–46, determine whether the statement is true or false. Justify your answer.**

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

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

45. Data that are modeled by

$y = 3.29 - 4.17x$

have a negative linear correlation.

46. When the correlation coefficient for a linear regression model is close to -1, the regression line is not a good fit for describe the data.

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

48. **Think About It** Construct a collection of numbers that has the following measures of central tendency.

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

Chapter Test

See CalcChat.com for tutorial help and worked-out solutions to odd-numbered exercises.

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 1–5, use the data set of the ages of all customers at an electronics store listed below.

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 for the data.

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

3. Construct a frequency distribution for the data.

4. Construct a histogram for the data.

5. You arrange the ages in increasing order, and then select the 4th, 8th, 12th, 16th, and 20th data value. Identify the type of sample you selected.

In Exercises 6 and 7, find the mean, median, and mode of the data set.

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

7. 25, 39, 36, 43, 37, 36

In Exercises 8 and 9, find the mean ($\bar{x}$), variance (v), and standard deviation (σ) of the data set.

8. 0, 1, 2, 3, 4, 5, 6

9. 2, 2, 4, 5, 7, 7

In Exercises 10 and 11, find the lower and upper quartiles and sketch a box-and-whisker plot for the data set.

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

11. 25, 20, 22, 28, 24, 28, 25, 19, 27, 29, 28, 21

In Exercises 12 and 13, 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.

12. Between 19 and 31

13. At most 22

In Exercises 14–16, consider the points (0, 2), (3, 4), and (4, 5).

14. Find the least squares regression line for the points. Use a graphing utility to verify your answer.

15. Find the sum of the squared differences using the linear model you found in Exercise 14.

16. Find the correlation coefficient r. How well does the linear model you found in Exercise 14 fit the data?

17. Find the least squares regression parabola for the points $(-2, 10)$, $(-1, 4)$, $(0, 2)$, and $(1, 2)$. Use a graphing utility to verify your answer.

18. The guayule plant 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 between 7 and 14 inches?

(c) 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 $\{x_1, x_2, . . . , x_n\}$ is

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

Proof

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

$$\begin{aligned}
\sigma &= \sqrt{v} && \text{Square root of variance} \\
&= \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdot \cdot \cdot + (x_n - \bar{x})^2}{n}} && \text{Substitute definition of variance.} \\
&= \sqrt{\frac{(x_1^2 - 2x_1\bar{x} + \bar{x}^2) + (x_2^2 - 2x_2\bar{x} + \bar{x}^2) + \cdot \cdot \cdot + (x_n^2 - 2x_n\bar{x} + \bar{x}^2)}{n}} && \text{Square binomials.} \\
&= \sqrt{\frac{(x_1^2 + x_2^2 + \cdot \cdot \cdot + x_n^2) - 2(x_1\bar{x} + x_2\bar{x} + \cdot \cdot \cdot + x_n\bar{x}) + \underbrace{(\bar{x}^2 + \bar{x}^2 + \cdot \cdot \cdot + \bar{x}^2)}_{n \text{ terms}}}{n}} && \text{Regroup terms in numerator.} \\
&= \sqrt{\frac{(x_1^2 + x_2^2 + \cdot \cdot \cdot + x_n^2)}{n} - \frac{2(x_1\bar{x} + x_2\bar{x} + \cdot \cdot \cdot + x_n\bar{x})}{n} + \frac{n\bar{x}^2}{n}} && \text{Rewrite radicand as three fractions.} \\
&= \sqrt{\frac{(x_1^2 + x_2^2 + \cdot \cdot \cdot + x_n^2)}{n} - \frac{2\bar{x}(x_1 + x_2 + \cdot \cdot \cdot + x_n)}{n} + \frac{n\bar{x}^2}{n}} && \text{Factor } \bar{x} \text{ out of numerator of second fraction.} \\
&= \sqrt{\frac{(x_1^2 + x_2^2 + \cdot \cdot \cdot + x_n^2)}{n} - 2\bar{x}^2 + \bar{x}^2} && \text{Definition of } \bar{x}\text{; divide out } n \text{ in third fraction.} \\
&= \sqrt{\frac{(x_1^2 + x_2^2 + \cdot \cdot \cdot + x_n^2)}{n} - \bar{x}^2} && \text{Simplify.}
\end{aligned}$$

■

P.S. Problem Solving

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

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 are listed below. *(Spreadsheet at LarsonPrecalculus.com)*

21, 34, 29, 38, 30, 54, 39, 20, 35, 64, 37, 45, 21, 25, 55, 45, 40, 38, 38, 21, 21, 34, 38, 52, 41, 48, 18, 27, 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.

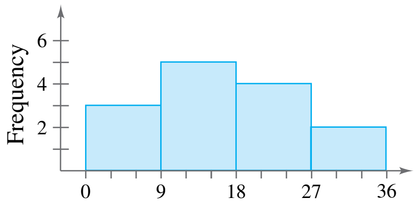
(c) According to one source, the age of passenger Thomas English was unknown at the time of the *Mayflower's* departure. What is the probability that he was 18–29 years old? Explain.

(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)	7	7
[25, 35)	11	7 + 11 = 18
?	?	?

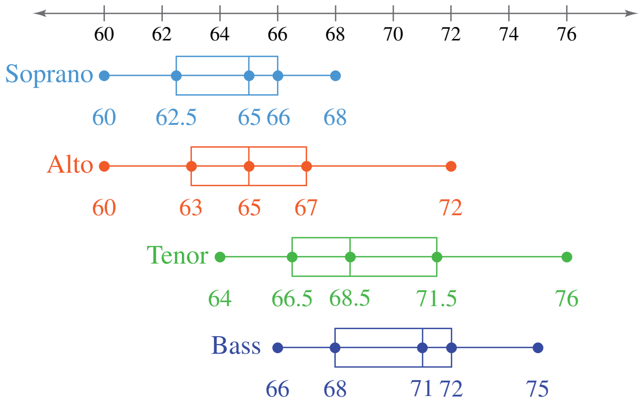
3. Organizing Data Construct a line plot, frequency distribution, and histogram for the data given in (a) Exercise 1 and (b) Exercise 2.

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. Describe the process you used.



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? If so, give an example.

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

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

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 approximately 69 inches and a standard deviation of approximately 3.1 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 of 235 and a standard deviation of 28. 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 if the standard deviation was 16?