# Probability and Statistics

# 6

## Unit Overview

In this unit you will investigate univariate data, using statistics and graphs to compare different distributions and to comment on similarities and differences among them. You will also use two-way tables to summarize bivariate categorical data and find a "best-fit line" to summarize bivariate numerical data. You will use technology to calculate a measure of strength and direction for relationships that are linear in form. Finally, you will learn to distinguish between correlation/association and causation.

## Key Terms

As you study this unit, add these and other terms to your math notebook. Include in your notes your prior knowledge of each word, as well as your experiences in using the word in different mathematical examples. If needed, ask for help in pronouncing new words and add information on pronunciation to your math notebook. It is important that you learn new terms and use them correctly in your class discussions and in your problem solutions.

### Academic Vocabulary

- cluster
- associate

### Math Terms

- sample
- sampling error
- measurement error
- standard deviation
- outlier
- normal distribution
- *z* score
- correlate
- correlation coefficient
- residual
- best-fit line
- segmented bar graph
- row percentages

## ESSENTIAL QUESTIONS

? How are dot plots, histograms, and box plots used to learn about distributions of numerical data?

? How can the scatter plot, best-fit line, and correlation coefficient be used to learn about linear relationships in bivariate numerical data?

? How can a two-way table be used to learn about associations between two categorical variables?

? When is it reasonable to interpret associations as evidence for causation?

## EMBEDDED ASSESSMENTS

These assessments, following Activities 37 and 40, will give you an opportunity to demonstrate your understanding of how graphical displays and numerical summaries are used to compare distributions and of methods for summarizing and describing relationships in bivariate data.

| | |
|---|---|
| **Embedded Assessment 1:** | |
| Comparing Univariate Distributions | p. 557 |
| **Embedded Assessment 2:** | |
| Bivariate Distributions | p. 609 |

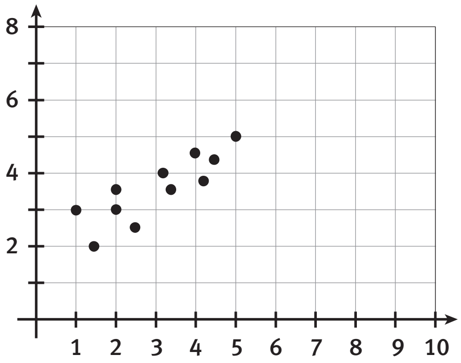
UNIT 6

# Getting Ready

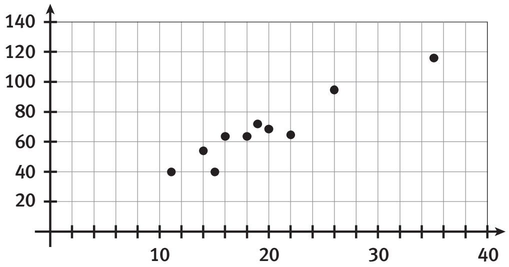
**Write your answers on notebook paper. Show your work.**

**1.** Each scatter plot below shows a set of $(x, y)$ coordinate pairs with an approximate linear trend. Estimate the equations of the trend lines for the graphs below.

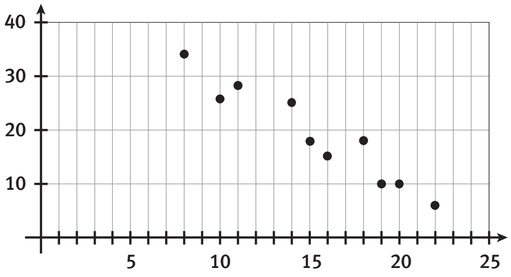
**a.**



**b.**



**c.**



**2.** For each trend line below, interpret the slope of the trend line in relation to the variable quantities.

**a.** $y = 75 + 0.19x$, where $x$ is the number of miles traveled and $y$ is the cost of an airline flight in dollars

**b.** $y = 100 - 1.2x$, where $x$ is the number of hours of TV watched per week and $y$ is the test score on last week's test

**3.** An election was held at Greg's school; Greg and his friend Mary were both nominated. The table below shows the results of the voting.

| Grade | Voted for Greg | Voted for Mary | Voted for Another Candidate | Total |
|---|---|---|---|---|
| Seventh | 63 | 81 | | 150 |
| Eighth | | 71 | 13 | 250 |
| Total | 229 | | 19 | 400 |

**a.** Fill in the three remaining cells of the table above.

**b.** What percentage of seventh graders voted for Greg? What percentage of eighth graders voted for Greg?

**4.** The heights (to the nearest inch) of 14 students are given below. Use these data for Parts (a)–(d).

| | | | | | | |
|---|---|---|---|---|---|---|
| 68 | 66 | 67 | 70 | 66 | 68 | 67 |
| 69 | 65 | 67 | 64 | 66 | 63 | 65 |

**a.** Compute the mean height of these students.

**b.** Compute the median height of these students.

**c.** Construct a dot plot of the heights of the students.

**d.** Describe the shape of the distribution shown in the dot plot.

# Measures of Center and Spread

ACTIVITY 36

## To Text, or Not to Text

## Lesson 36-1 Mean, Median, Mode, and MAD

My Notes

### Learning Targets:

- Interpret differences in center and spread of data in context.
- Compare center and spread of two or more data sets.
- Determine the mean absolute deviation of a set of data.

**SUGGESTED LEARNING STRATEGIES:** Summarizing, Interactive Word Wall, Create Representations, Look for a Pattern, Think-Pair-Share

Zach is a high school student who enjoys texting with friends after school. Recently, Zach's parents have become concerned about the amount of time that he spends text messaging on school nights.

Zach decides to compare the amount of time he spends text messaging to that of his good friend Olivia. Both of them record the number of minutes they spend text messaging on school nights for one week.

| | Sunday | Monday | Tuesday | Wednesday | Thursday |
|---|---|---|---|---|---|
| **Zach** | 10 min | 60 min | 20 min | 135 min | 75 min |
| **Olivia** | 60 min | 60 min | 60 min | 60 min | 60 min |

One way to describe a set of data is by explaining how the data ***cluster*** around a value, or its center. The measures of center include the **mean,** the **median,** and the **mode**.

**ACADEMIC VOCABULARY**

To ***cluster*** means to group around.

**1.** Find the mean amount of time that Zach spends text messaging each night. Show how you determined your answer.

**2.** Find the mean amount of time that Olivia spends text messaging each night. Show how you determined your answer.

My Notes

3. **Reason quantitatively.** Compare the amounts of time that Zach and Olivia spend text messaging. Describe similarities and differences.

### MATH TIP

Another word for spread is **variability.**

Zach knows that data can be described by center and also by spread. **Spread** indicates how far apart the data values are in the set. Measures of spread include the **range** and the **mean absolute deviation**.

Zach asks his friend Trey to record the amount of time he spends text messaging on school nights. To measure spread, Zach chooses the *range*.

4. Find the mean and range of Trey's data.

| | Sunday | Monday | Tuesday | Wednesday | Thursday |
|---|---|---|---|---|---|
| Zach | 10 min | 60 min | 20 min | 135 min | 75 min |
| Trey | 10 min | 10 min | 135 min | 135 min | 10 min |

5. Complete the table below. How do the mean and range of Trey's data compare to those of Zach's?

| | Mean | Range |
|---|---|---|
| Trey | | |
| Zach | | |

6. **Construct viable arguments.** Describe how the two data sets are different. Did the mean and range help you to identify these differences? Explain.

## Lesson 36-1

### Mean, Median, Mode, and MAD

My Notes

Because the range is based on only two values, it does not reflect any variation in the data between the greatest and least values. The range is greatly influenced by extreme values.

Another measure of spread that is not as influenced by extremes is the mean absolute deviation, which is computed using all the data values. The **mean absolute deviation** is the mean (average) of the absolute values of the deviations of the data. The **deviation** is a measure of how far a data value is from the mean.

**7.** To find the mean absolute deviation of Zach's data, begin by completing the table. Use the mean for Zach's data that you calculated in Item 1.

| Zach | Deviation | Absolute Deviation |
|---|---|---|
| Time<br>$x$ | Time − Mean<br>$= (x - \overline{x})$ | $\lvert$Time − Mean$\rvert$<br>$= \lvert x - \overline{x} \rvert$ |
| 10 | | |
| 60 | | |
| 20 | | |
| 135 | | |
| 75 | | |

**WRITING MATH**

The symbol $\overline{x}$ is used to represent the mean of a set of values.

**8.** To finish calculating the mean absolute deviation of Zach's data, find the mean of the numbers in the third column. Determine the sum of the numbers in the third column and then divide by the number of data values (the number of items in the first column).

My Notes

**9.** **Reason abstractly.** Why would statisticians use the mean absolute deviation rather than the mean of the deviations (in the second column)?

| | Sunday | Monday | Tuesday | Wednesday | Thursday |
|---|---|---|---|---|---|
| **Zach** | 10 min | 60 min | 20 min | 135 min | 75 min |
| **Olivia** | 60 min | 60 min | 60 min | 60 min | 60 min |
| **Trey** | 10 min | 10 min | 135 min | 135 min | 10 min |

**10.** Trey's text messaging minutes are shown in the table above.

**a.** Find the mean absolute deviation for the amount of time that Trey spends text messaging.

**b.** Why is the mean absolute deviation for Trey's data set greater than the mean absolute deviation for Zach's data set?

**11.** Olivia's text messaging minutes are also given in the table above Item 10.

**a.** Find the mean absolute deviation for the amount of time that Olivia spends text messaging.

**b.** **Make sense of problems.** Explain why Olivia's mean absolute deviation is descriptive of her data.

My Notes

## Check Your Understanding

**12.** During the annual food drive, Mr. Binford's homeroom collected canned goods for a month. The numbers of cans collected are given below.

Boys

| | | | | | |
|---|---|---|---|---|---|
| 12 | 42 | 69 | 91 | 97 | 61 |
| 15 | 37 | 104 | 38 | 82 | 90 |
| 51 | 96 | 19 | 66 | 8 | 24 |

Girls

| | | | | | |
|---|---|---|---|---|---|
| 20 | 63 | 18 | 89 | 67 | 19 |
| 66 | 108 | 96 | 24 | 16 | 44 |

Compare and contrast the results for the boys and girls using the mean and the range.

**13.** Remi recorded data on her car's fuel efficiency for five trips in the table below.

| **Trip** | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| **Miles per Gallon** | 23.7 | 25.5 | 25.2 | 24.8 | 25.4 |

**a.** Calculate the absolute deviation for each trip if the average number of miles per gallon for the trips was 24.9.

**b.** Find the mean absolute deviation for the five trips.

Zach would like to assure his parents that his text messaging time is not unusual for a high school student. His average text messaging time is the same as Olivia's and Trey's average times, but the variability differs for each set of data.

Before reporting to his parents, Zach decides to gather additional information. He considers using a ***census*** of the 1250 students in his school.

**14.** Conducting a census is often difficult. What difficulties might Zach encounter if he proceeds with his census?

## ACADEMIC VOCABULARY

A ***census*** gathers information about every member of the population.

My Notes

## MATH TERMS

A **sample** is a portion of the population. A good sample looks like the entire population and provides useful information about the population.

**Sampling error** occurs because particular subgroups of the population are missing from the sample. Sampling error is also called **sample selection bias**.

Instead of gathering information from all 1250 students in the school, Olivia suggests that Zach conduct a survey of 40 students.

**15.** What is the population of Zach's survey? What is the **sample**?

Olivia warns Zach to choose his sample wisely. A selection method that produces samples that are not representative of the total population will introduce a **sampling error** into the process.

**16.** Explain why each sampling method produces sampling error when surveying the school's population.

**a.** Zach asks the 30 students in his algebra class how many minutes they spend per school night text messaging with friends.

**b.** Zach leaves questionnaires on a table in the main hall. One of the questions asks students to indicate the number of minutes they spend per school night text messaging with friends.

Only sampling methods that incorporate random chance into the sample selection method can hope to avoid sampling error.

**17.** Zach obtains a list of the names of every student in the school. Explain how Zach could use the list to randomly select a sample of 40 students.

**18.** Error can also occur when the method for obtaining a response is flawed. Suppose that Zach has properly selected a random sample of 40 students. Determine why each method produces **measurement error**.

**a.** Zach gives each student a questionnaire. One question states: "It is very important for young people to have time to socialize with each other and today's method of communication seems to be text messaging. How many minutes do you typically spend text messaging with friends on school nights?"

**b.** Zach verbally surveys 40 students. Zach feels that the student responses are too low, so he reminds them that he is trying to convince his parents that 60 minutes per night is not too much time to spend text messaging.

My Notes

### MATH TERMS

**Measurement error** occurs when incorrect or misleading data are collected that can be ascribed to the interviewer, the respondent, the survey instrument, or the method used for recording the data.

### CONNECT TO AP

In AP Statistics, it is important to understand whether a particular sample or method for gathering information contains the potential for error or bias.

My Notes

**19.** **Attend to precision.** Gathering good data involves avoiding error in the process.

**a.** Write a question that Zach can use to gather data about text messaging habits and avoid measurement error.

**b.** Describe a method that Zach can use to collect answers from a group of 40 students and avoid sampling error.

## Check Your Understanding

Classify each as a sample or a census.

**20.** Surveying every ninth-grade student regarding the new dress code policy for the students in all grades in the school to determine the student opinion

**21.** Asking every eighth-grade student whom they will vote for in the upcoming eighth-grade student election

**22.** Selecting every fourth student from an alphabetical list of students to gather data on absenteeism for the school

## LESSON 36-1 PRACTICE

My Notes

**23.** Vernice asked 12 classmates to record the number of hours they spent watching television during one week. The table shows the data she collected.

| | | | | | |
|---|---|---|---|---|---|
| 10 | 11 | 22 | 7 | 17 | 17 |
| 20 | 31 | 0 | 12 | 19 | 23 |

**a.** Calculate the mean and mean absolute deviation for the data.

**b.** What statements could Vernice make about the viewing habits of these classmates?

**24.** Rewrite each question to avoid measurement error.

**a.** Don't you agree that seniors should be dismissed early on Fridays at least once each semester?

**b.** Drinking beverages with sugar promotes tooth decay and obesity. How many soft drinks with sugar did you drink in the past week?

**25.** Scores from the same benchmark test were collected from two algebra classes, each with 30 students enrolled. One class had a mean score of 79 with a mean absolute deviation of 5, and the other had a mean score of 81 with a mean absolute deviation of 10. What can be said about the distribution of scores on this test for the two classes?

**26.** Mitch and a group of his friends have estimated how long it will take each of them to run 400 meters around the track. Their estimates in seconds are 115, 76, 94, 81, 78, 99, 68, and 84.

**a.** Calculate the mean and the range.

**b.** Which estimate stands out as unusual?

**c.** What might be a reason for such an unusual estimate?

**27.** **Construct viable arguments.** What are the advantages of taking a census of the population instead of a sample? Describe some examples when it would be worth the time and effort to conduct a census.

My Notes

## Learning Targets:

- Use summation and subscript notation.
- Calculate and interpret the standard deviation of a numerical data set.
- Select appropriate measures of spread by examining the shape of a distribution.

**SUGGESTED LEARNING STRATEGIES:** Summarizing, KWL Chart, Create Representations, Self Revision/Peer Revision, Think-Pair-Share

In the previous lesson, you studied one way to describe the spread, or variability, in a data set—the mean absolute deviation (MAD). The mean absolute deviation is the mean (or average) difference of the data values from the mean of a numerical data set.

Consider the following data from C|NET (www.cnet.com), a tech media website that publishes information about technology and consumer electronics. These data are taken from a review of cell phone battery lifetimes. The table below gives the talk times (in hours) for the top 10 brands of batteries.

| Talk Time (hours) | |
|---|---|
| 19.78 | 11 |
| 14.55 | 10.7 |
| 13.4 | 10.6 |
| 12.75 | 10.6 |
| 12 | 10.3 |

**1.** Calculate the mean for these 10 talk times.

**2.** Complete the table below by finding the absolute values of the deviations (the absolute value of the difference between each data value and the mean calculated in Item 1).

| Talk Time and Deviations from the Mean | | | |
|---|---|---|---|
| **Value** | **\|Deviation\|** | **Value** | **\|Deviation\|** |
| 19.78 | 7.212 | 11.00 | 1.568 |
| 14.55 | | 10.70 | |
| 13.40 | | 10.60 | |
| 12.75 | | 10.60 | |
| 12.00 | | 10.30 | |

## Lesson 36-2
### Another Measure of Variability

My Notes

**3.** **Attend to precision.** The mean absolute deviation for this data set is the mean of the 10 absolute deviations. Calculate the mean absolute deviation.

Another common measure of variability is the **standard deviation**. Before calculating the standard deviation, let's introduce some notation.

### MATH TERMS

The **standard deviation** is a measure of variability in a data set.

The Greek letter $\sum$ (sigma) is used to indicate a sum of several values.

For example, $\sum_{i=1}^{5} i$ means "the sum of the values of *i* as *i* goes from 1 to 5":

*i* ends at this value.

$$\sum_{i=1}^{5} i = 1 + 2 + 3 + 4 + 5 = 15$$

*i* starts at this value.

In the summation above, *i* can be replaced by any expression. Also, the values of *i* can start and end at any number:

$$\sum_{i=3}^{6} (i + 2) = (3 + 2) + (4 + 2) + (5 + 2) + (6 + 2) = 26$$

### MATH TIP

In sigma notation, the letter *i* is called the **index of summation.**

**4.** Determine each sum.

**a.** $\sum_{i=1}^{4} i^2$

**b.** $\sum_{i=0}^{2} 2^i$

**c.** $\sum_{i=2}^{7} |5 - i|$

The standard deviation is similar to the MAD in that it is based on deviations from the mean. The formulas below show the similarities between the MAD and the standard deviation *s*.

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

### MATH TIP

In a data set with *n* values, the individual values are referred to as $x_1, x_2, x_3, ..., x_n$. In the formulas for the MAD and the standard deviation, $x_i$, as *i* goes from 1 to *n*, refers to these values.

As you can see, the calculation of the standard deviation is a little more involved than MAD. Why this additional complexity? The basic answer is that this measure has some advantages in more advanced statistical settings.

You will calculate both the MAD and standard deviation using a data set consisting of four battery talk times (in hours):

$$x_1 = 7.00,\ x_2 = 10.00,\ x_3 = 8.10,\ x_4 = 9.32$$

The mean of these 4 observations is 8.605.

My Notes

**5.** Complete the table. Then compute the MAD and the standard deviation.

| $x_i$ | $\bar{x}$ | $\lvert x_i - \bar{x}\rvert$ | $(x_i - \bar{x})^2$ |
|---|---|---|---|
| $x_1 = 7.00$ | 8.605 | | |
| $x_2 = 10.00$ | 8.605 | | |
| $x_3 = 8.10$ | 8.605 | | |
| $x_4 = 9.32$ | 8.605 | | |
| $\sum_{i=1}^{4} x_i =$ | | $\sum_{i=1}^{4}\lvert x_i - \bar{x}\rvert =$ | $\sum_{i=1}^{4}(x_i - \bar{x})^2 =$ |

$$\text{mean absolute deviation} = \frac{\sum_{i=1}^{4}\lvert x_i - \bar{x}\rvert}{n} =$$

$$\text{standard deviation} = \sqrt{\frac{\sum_{i=1}^{4}(x_i - \bar{x})^2}{n-1}} =$$

**6.** The weights (in ounces) of three newborns are:

$$w_1 = 120; \quad w_2 = 115; \quad w_3 = 125$$

**a.** Compute the mean of these three weights.

**b.** Complete the table below and calculate both the MAD and the standard deviation.

| $w_i$ | $\bar{w}$ | $\lvert w_i - \bar{w}\rvert$ | $(w_i - \bar{w})^2$ |
|---|---|---|---|
| 120 | | | |
| 115 | | | |
| 125 | | | |
| $\sum_{i=1}^{3} w_i =$ | | $\sum_{i=1}^{3}\lvert w_i - \bar{w}\rvert =$ | $\sum_{i=1}^{3}(w_i - \bar{w})^2 =$ |

$$\text{mean absolute deviation} = \frac{\sum_{i=1}^{3}\lvert w_i - \bar{w}\rvert}{n} =$$

$$\text{standard deviation} = \sqrt{\frac{\sum_{i=1}^{3}(w_i - \bar{w})^2}{n-1}} =$$

You can see that calculating the standard deviation can be a lot of work. Fortunately, calculators and computers can be used to do the calculations.

My Notes

## Check Your Understanding

**7.** Calculate the mean and standard deviation of the original data for the cell phone battery talk times (shown below) using a calculator or computer software.

| Talk Times (hours) | |
|---|---|
| 19.78 | 11.00 |
| 14.55 | 10.70 |
| 13.40 | 10.60 |
| 12.75 | 10.60 |
| 12.00 | 10.30 |

## LESSON 36-2 PRACTICE

The table shows the speeds of the 10 fastest roller coasters in the United States. Use the table for Items 8–11.

| Fastest Roller Coasters (mi/h) | |
|---|---|
| 76 | 81 |
| 85 | 67 |
| 74 | 72 |
| 63 | 59 |
| 73 | 80 |

**8.** Find $|x - \bar{x}|$ for each data value.

| **Value** | 76 | 81 | 85 | 67 | 74 | 72 | 63 | 59 | 73 | 80 |
|---|---|---|---|---|---|---|---|---|---|---|
| $\lvert x - \bar{x} \rvert$ | | | | | | | | | | |

**9.** Find the mean absolute deviation using the values you calculated in Item 8.

**10.** Using technology, calculate the standard deviation for the data set.

**11.** **Make sense of problems.** A new roller coaster is being built, scheduled to be complete in the spring of next year. It is projected to reach speeds of 90 mi/h. This new value will cause the lowest speed to be removed from the list. Calculate the new standard deviation and describe the changes.

## ACTIVITY 36 PRACTICE

**Write your answers on notebook paper.
Show your work.**

**1.** For each question, decide:
- Is it a statistical question?
- If not, explain why not and rewrite the question so that it is a statistical question.

**a.** How many states have you visited?

**b.** Do you like chocolate ice cream?

**c.** Do you watch TV at night?

**d.** How many sports do you play?

**2.** Write three examples of statistical questions whose responses would show varying levels of variability. Include at least one question whose responses would show "lots" of variability and at least one question whose responses would show "little" variability.

**3.** Describe the variability associated with the question "Do students in my school need more homework each night?" Do you think there will be a lot or a little variability? Explain.

**4.** Suppose the results of a survey about where your classmates were born showed great variability. What would these results tell you about the people in your class?

**5.** The following question was asked of your classmates today before school started:

"How many states have you visited in the last year?"

The responses to this question showed little variability. How could you change this question to gather answers that show greater variability?

For Items 6 and 7, use the data set $x_1 = 2.3,\ x_2 = 4.1,\ x_3 = 1.6,$ and $x_4 = 2.0.$

**6.** Compute the value of $\frac{x_1 + x_2}{x_3}$.

**7.** Compute the value of $\sum_{i=1}^{3} x_i$.

Use the information below for Items 8 and 9.

The amount of caffeine in beverages presents an important health concern, especially for women of childbearing age. In a recent study of carbonated sodas, the numbers of milligrams of caffeine detected were as follows:

29.5, 38.2, 39.6, 29.5, 31.7, 27.4, 45.4, 48.2, 36.0, 33.8, 19.4, 18.0, 34.6

**8.** Calculate the mean, mean absolute deviation, and standard deviation for these data.

## MATHEMATICAL PRACTICES
### Reason Abstractly and Quantitatively

**9.** In the report of the caffeine levels, five sodas were left out because no caffeine was detected. If these sodas were given a value of 0 mg of caffeine and added to the data above, how would the mean and standard deviation change?

# Dot and Box Plots and the Normal Distribution

ACTIVITY 37

## Disturbing Coyotes

## Lesson 37-1 Dot Plots and Box Plots

My Notes

### Learning Targets:

- Construct representations of univariate data in a real-world context.
- Describe characteristics of a data distribution, such as center, shape, and spread, using graphs and numerical summaries.
- Compare distributions, commenting on similarities and differences among them.

**SUGGESTED LEARNING STRATEGIES:** Summarizing, Paraphrasing, Look for a Pattern, Discussion Groups, Quickwrite

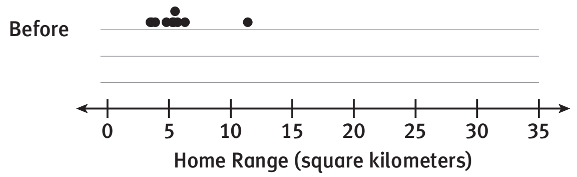
Professional wildlife managers and the public are concerned with the impact of human activity on wildlife. One measure studied is animals' "home range," the typical area in which an animal spends its time.

Researchers were concerned that the home ranges of some coyotes in a portion of Colorado were affected by military maneuvers involving jeeps, tanks, helicopters, and jet fighter flyovers. To evaluate these potential effects, several coyotes were collared with radio transmitters. The researchers used the transmitters to track the movement of the coyotes. Coyotes were monitored before, during, and after the military maneuvers.

| Home Range Before Maneuvers ($km^2$) | Home Range During Maneuvers ($km^2$) | Home Range After Maneuvers ($km^2$) |
|---|---|---|
| 3.9 | 7.5 | 3.1 |
| 5.4 | 32.6 | 5.4 |
| 5.7 | 3.2 | 4.5 |
| 4.8 | 2.7 | 4.1 |
| 5.3 | 7.3 | 8.6 |
| 5.3 | 9.1 | 8.0 |
| 5.5 | 18.0 | 6.4 |
| 11.4 | 6.5 | 7.8 |
| 3.6 | 2.1 | 1.0 |
| 3.5 | 5.2 | 3.7 |
| 6.3 | 4.3 | 10.7 |

A **dot plot** is an effective method for representing univariate (one-variable) data when dealing with small data sets.

1. A dot plot of the "Before" data is shown below. Make dot plots of the "During" and "After" data sets.



**MATH TIP**

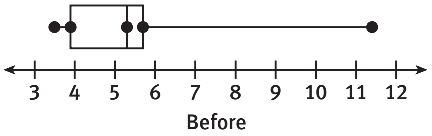
To make a dot plot, draw a number line with an appropriate scale for the data. Then mark a dot for each data value above the appropriate number on the number line.

My Notes

**2.** Compare and contrast the centers and spreads of the three data sets.

A five-number summary provides a numerical summary of a set of data. It is used to construct a **box plot.** The five-number summary for the "Before" home ranges is shown below, together with the resulting box plot.

| Home Ranges: Before Maneuvers | |
|---|---|
| Minimum: | 3.5 |
| First quartile ($Q_1$): | 3.9 |
| Median: | 5.3 |
| Third quartile ($Q_3$): | 5.7 |
| Maximum: | 11.4 |



### MATH TIP

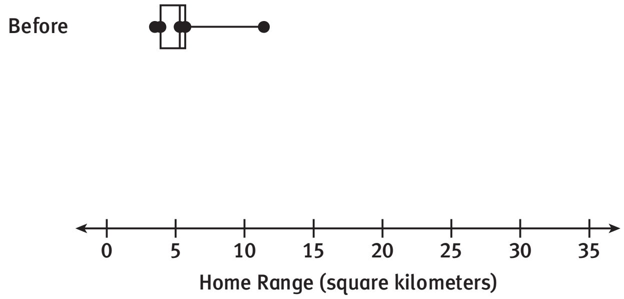
The first quartile ($Q_1$) is the median of the data values to the left of the overall median, and the third quartile ($Q_3$) is the median of the data values to the right of the overall median. A box plot (sometimes called a box-and-whisker plot) is a graph of the five-number summary that consists of a central box from $Q_1$ to $Q_3$ that has a vertical line segment at the median. Horizontal line segments ("whiskers") extend from the box to the minimum and maximum data values.

**3.** Create five-number summaries of the "During" and "After" data.

| Home Ranges: During Maneuvers | |
|---|---|
| Minimum: | |
| First quartile: | |
| Median: | |
| Third quartile: | |
| Maximum: | |

| Home Ranges: After Maneuvers | |
|---|---|
| Minimum: | |
| First quartile: | |
| Median: | |
| Third quartile: | |
| Maximum: | |

**4.** Use the summaries in Item 3 to construct box plots of the "During" and "After" data sets in the space below.



**5.** Based on the box plots and five-number summaries in Items 3 and 4:

**a.** Which data set seems to have the least overall spread? Which data set seems to have the greatest overall spread?

**b.** Which data set seems to have the least spread in its "middle 50%" box? Which data set seems to have the greatest spread in its "middle 50%" box?

Remember that the initial concern before the data gathering was that the home ranges of the local coyotes might change during the military maneuvers. To investigate this concern, you will use the graphs, numerical summaries, and comparisons you have developed as a starting point for your analysis.

**6.** **Reason abstractly and quantitatively.** Based on the data and graphs, does it appear that there was a substantial change in the coyotes' home ranges during the military maneuvers? Write a few sentences specifically comparing the "Before" and "During" data sets. Use numerical values where possible.

**7.** Do there appear to be any substantial permanent changes to the coyotes' home ranges after the military maneuvers? Write a few sentences specifically comparing the "Before" and "After" data sets. Use numerical values where possible.

**8.** The “During” data set contains two values that are far away from the rest of the data. The “Before” data set contains one such value also. Suppose you wish to call attention to the fact there are such far-away data values. Which type of plot—the dot plot or the box plot—would be your choice? Why?

## Check Your Understanding

**9.** A teacher in a statistics class allows her students to use notes about statistical procedures on tests. She believes that a teacher-made study sheet will be more effective in helping students recall the procedures. In each of her three classes she used one of three helping strategies: (a) student-made notes with information about the procedures, (b) teacher-made information printed on paper in the form of a flowchart, and (c) teacher-made information delivered by computer access during the exam. Each of her classes has 18 students. The test scores for her students are given as percent correct.

| | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Student Notes** | 89 | 15 | 39 | 15 | 31 | 69 | 39 | 54 | 31 | 62 | 46 | 39 | 54 | 39 | 15 | 46 | 23 | 31 |
| **Paper Help** | 76 | 24 | 77 | 71 | 18 | 29 | 59 | 77 | 41 | 77 | 77 | 47 | 71 | 82 | 82 | 82 | 59 | 65 |
| **Computer Help** | 100 | 13 | 73 | 73 | 33 | 53 | 60 | 60 | 27 | 80 | 80 | 47 | 73 | 80 | 80 | 93 | 60 | 53 |

**a.** In order to compare the results for these three groups, construct a dot plot for each of the three data sets.

**b.** Describe the three data sets with specific attention to center and spread.

**10.** For each group, create a five-number summary for these data.

| | Student Notes | Paper Help | Computer Help |
|---|---|---|---|
| Minimum | | | |
| First quartile | | | |
| Median | | | |
| Third quartile | | | |
| Maximum | | | |

**11.** Use the summaries in Item 10 to construct box plots of the three data sets: “Notes,” “Paper,” and “Computer.”

My Notes

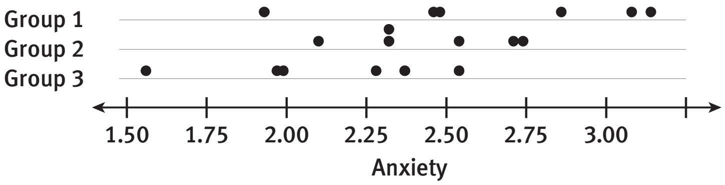
**12.** **a.** For these data, what advantages do you see in using the dot plots to display the data sets?
**b.** For these data, what advantages do you see in using the box plots to display the data sets?

**13.** Comparing the test scores for these groups and using specific information from the five-number summary and/or your dot plots and box plots, answer the following in a few sentences.
**a.** Which group appears to have done the best on the exam?
**b.** Which group appears to have done the worst on the exam?

## LESSON 37-1 PRACTICE

Some researchers believed that one reason students often have unhealthy sleeping habits is that they don't adequately manage their time. The researchers wanted to test whether providing information to students about time management could help. Eighteen students were divided into three groups. Students in Group 1 were taught to use a planner for time management and asked to sleep 7–8 hours daily. Students in Group 2 were taught to use a planner but not given any instruction on how many hours to sleep daily. Students in Group 3 were simply instructed to sleep as they usually do. At the conclusion of the study the participants were given a questionnaire to measure their anxiety levels. (A high score on the questionnaire indicates low levels of anxiety.)

Dot plots of the data from the questionnaire are shown below. Use the dot plots for Items 14 and 15.



**14.** In a few sentences, compare the centers of these data sets. Do the three groups have approximately equal centers? If not, how do they differ?

**15.** In a few sentences, compare the spreads of these data sets. Do the three groups have very similar spreads? If not, how do they differ?

My Notes

A table of the data from the questionnaires is shown below. Use the table for Item 16.

| Group 1: Planner & Sleep Instruction | Group 2: Planner Only | Group 3: Neither Planner nor Sleep Instruction |
|---|---|---|
| 2.86 | 2.32 | 1.56 |
| 3.14 | 2.71 | 1.97 |
| 2.48 | 2.32 | 2.28 |
| 1.93 | 2.54 | 2.54 |
| 2.46 | 2.74 | 1.99 |
| 3.08 | 2.10 | 2.37 |

**16.** Create a five-number summary for each group.

| | Group 1: Planner & Sleep Instruction | Group 2: Planner Only | Group 3: Neither Planner nor Sleep Instruction |
|---|---|---|---|
| Minimum | 1.93 | 2.10 | 1.56 |
| First quartile | | 2.32 | |
| Median | | 2.43 | 2.135 |
| Third quartile | | | 2.37 |
| Maximum | 3.14 | 2.74 | 2.54 |

**17.** Use the five-number summaries you created in Item 16 to draw box plots for the three groups.

**18.** **Use appropriate tools strategically.** You now have two different graphs for these data. Which graph—dot plots or box plots—do you feel makes it easier to compare centers and spreads? Explain your reasoning.

My Notes

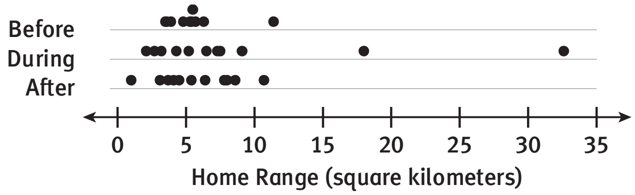
### Learning Targets:

- Use modified box plots to summarize data in a way that shows outliers.
- Compare distributions, commenting on similarities and differences among them.

**SUGGESTED LEARNING STRATEGIES:** Summarizing, Paraphrasing, Think-Pair-Share, Create Representations, Quickwrite

You already are familiar with many ways to summarize data graphically and numerically. In this lesson, you will see a new type of plot called a **modified box plot.**

Below are the dot plots of the data about the coyotes from Lesson 37-1. Refer back to Lesson 37-1 for the actual data values.



Notice that there are two unusually large values in the "During" data set and one in the "Before" data set.

**Outliers** are values that differ so much from the rest of a one-variable data set that attention is drawn to them. Outliers may arise for many reasons, including measurement errors or recording errors. It may also be the case that there actually are unusual values in a data set.

Outliers can also occur in bivariate (two-variable) data. It is important to consider the impact of outliers when summarizing and analyzing data.

**MATH TERMS**

An **outlier** is a data point that is unusual enough that it draws attention during the data analysis.

**1.** Which, if any, of the data values shown in the dot plots do you consider to be outliers? List each data value that you think might be an outlier and the data set from which it came.

**2.** Describe the method you used in Item 1 to determine whether a data value is an outlier. Is it based on distance? Is it based on the concentration of other data values? Compare your method with those of your classmates.

My Notes

## MATH TIP

**Notation Review**

$Q_1$ First quartile

$Q_3$ Third quartile

$IQR = Q_3 - Q_1$

It is not surprising that people do not always agree about how different a value must be in order to be called an outlier. For consistency, a data value is considered to be an outlier if it is more than 1.5 × $(IQR)$ from the nearest quartile. Recall that $IQR$ stands for **interquartile range** and is the difference between the third quartile $(Q_3)$ and the first quartile $(Q_1)$: $IQR = Q_3 - Q_1$.

Mathematically, this means that a data value $x$ is an outlier if:

$$x > Q_3 + 1.5(IQR)$$
or
$$x < Q_1 - 1.5(IQR)$$

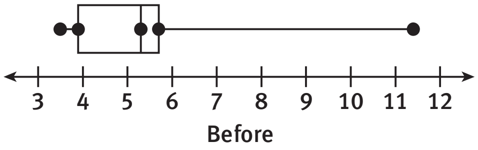
We will use this definition of an outlier to draw a *modified* box plot. The idea behind modifying the box plot is to create a plot that shows outliers.

When creating a *modified* box plot, the procedure for determining the length of the whiskers is different. The whiskers extend only to the least data value that is not an outlier and to the greatest data value that is not an outlier. Outliers are then shown by adding dots to the box plot to indicate their locations.

Let's walk through the steps together using the data from the "Before" data set. Here are the data, arranged in order, as well as the five-number summary and box plot of the data from Lesson 37-1:

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 3.5 | 3.6 | 3.9 | 4.8 | 5.3 | 5.3 | 5.4 | 5.5 | 5.7 | 6.3 | 11.4 |

| Home Ranges: Before Maneuvers | |
|---|---|
| Minimum: | 3.5 |
| First quartile: | 3.9 |
| Median: | 5.3 |
| Third quartile: | 5.7 |
| Maximum: | 11.4 |



The maximum value, 11.4, is much greater than the other data values. In fact, it is more than 5 units away from the next-greatest value of 6.3. The remaining data (from the minimum value of 3.5 to 6.3) have a range of only 2.8.

Now let's modify the box plot to show outliers.

My Notes

**3.** For the "Before" data set, how great or how small must a data value be to be identified as an outlier? Perform the calculations below to find the upper and lower boundary values that separate outliers from the rest of the data.

$$Q_3 + 1.5(IQR) =$$

$$Q_1 - 1.5(IQR) =$$

**4.** Are any data values less than the lower boundary for outliers identified in Item 3?

**5.** Are any data values greater than the upper boundary for outliers identified in Item 3?

Next, identify the least and greatest values in the data set that are not outliers. These values will determine the endpoints of the whiskers in the modified box plot.

**6.** What is the least data value that is not an outlier?

**7.** What is the greatest data value that is not an outlier?

My Notes

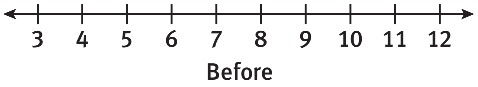
Now you have everything you need to draw the modified box plot. The modified box plot is constructed as follows:

**Step 1.** Draw the box as usual.

**Step 2.** Extend the whiskers to the least and greatest data values that are not outliers.

**Step 3.** Place dots above the scale to indicate the outliers.

**8.** Follow the directions above to draw the modified box plot for the "Before" data:



## Check Your Understanding

**9.** A data set has a third quartile of 64 and a first quartile of 29. What are the upper and lower boundary values that separate outliers from the rest of the data set?

**10. Make sense of problems.** If the third quartile of the data set in Item 9 were increased by 10, how would this change the upper boundary for outliers? Explain your answer.

**11. Critique the reasoning of others.** In a survey of 21 teenage girls about their text message usage for one month, the five-number summary is minimum = 0, $Q_1 = 1$, median = 31, $Q_3 = 56$, and maximum = 1305. In analyzing these data, Michelle determined that there are no outliers. Do you agree or disagree? Explain.

My Notes

## LESSON 37-2 PRACTICE

**12.** Describe the effects an outlier can have on a set of data.

**13. Construct viable arguments.** In a data set, the third quartile is 36 and the first quartile is 12. Would a value of 52 be considered an outlier? Why or why not?

**14.** Lisa recorded the heights of her classmates. Her data are shown in the table below.

| Heights of Classmates in Inches | | | | |
|---|---|---|---|---|
| 61 | 58 | 62 | 60 | 57 |
| 67 | 68 | 61 | 64 | 70 |
| 72 | 64 | 63 | 59 | 69 |

Calculate the upper and lower boundaries for outliers for the heights of Lisa's classmates.

**15.** List two values that would be considered outliers in the data set in Item 14. Include one value that is less than the lower boundary for outliers and one that is greater than the upper boundary for outliers.

My Notes

## Learning Targets:

- Use the mean and standard deviation to fit a normal distribution.
- Develop an understanding of the normal distribution.
- Use technology to estimate the percentages under the normal curve.

**SUGGESTED LEARNING STRATEGIES:** Visualization, Think-Pair-Share, Create Representations, Look for a Pattern, Quickwrite

There is a relationship between sleep quality and health. Many studies have concluded that good-quality rest helps to relieve health issues such as high blood pressure, depression, weight gain or loss, and fatigue.

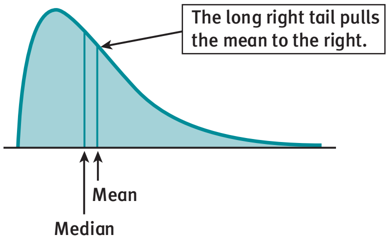
**1.** How many hours did you spend sleeping in the last 24 hours?

**2.** Gather all of your classmates' answers to Item 1. In the *My Notes* section of this page, create a dot plot of the number of hours that students in your class spent sleeping.

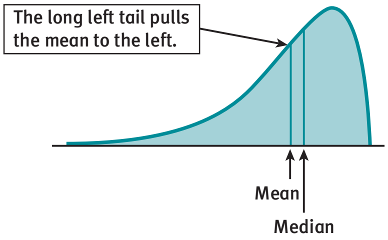
**3.** Use the dot plot to describe the data. Identify the mean, standard deviation, maximum, and minimum.

**4.** What do you consider to be a normal amount of time spent sleeping in a 24-hour period?

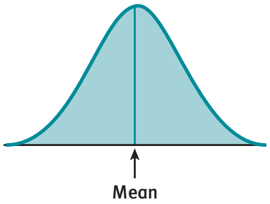
Data can be distributed in different ways.



The majority of the data can be grouped to the left (skewed right).



The majority of the data can be grouped to the right (skewed left).



The majority of the data can be grouped symmetrically around the center.

This last type of distribution is called a **normal distribution**.

### MATH TERMS

In a **normal distribution**, the data values are grouped symmetrically about the mean, with most of the data values occurring near the mean. Because of its shape, a normal distribution is sometimes called a bell curve.

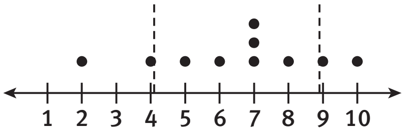
# Lesson 37-3
## Normally Distributed

Many real-world data are approximately normally distributed—for instance, height, weight, grades, blood pressure, and time people spend sleeping. Because so many data sets can be modeled by a normal distribution, a table of values is used to analyze and make decisions about normally distributed data. For this to be accomplished, the data values are converted to "scores" that can be easily compared. This is called *standardizing*. A standard score means that the data have a mean of 0 and a standard deviation of 1.

**5.** Below is a list of the hours slept by the students in Mr. Trent's class. Calculate the mean number of hours these students slept.
2, 4, 5, 6, 7, 7, 7, 8, 9, 10

**6.** Find the probability that a student in Mr. Trent's class slept more than 7 hours.

Below is a graph of the hours slept by Mr. Trent's students. Dashed vertical lines have been drawn at one standard deviation above and below the mean.



**7.** Calculate the percent of students whose time spent sleeping is within one standard deviation above or below the mean.

The first step in standardizing the data values from Mr. Trent's class is to calculate each value's deviation from the mean.

**8.** For each data value, calculate the deviation from the mean, $(x - \bar{x})$, and fill in the second column of the table. Leave the third column blank for now.

**Hours Slept and Deviation from the Mean**

| Hours | $(x - \bar{x})$ | $\frac{x - \bar{x}}{s}$ |
|---|---|---|
| 2 | | |
| 4 | | |
| 5 | | |
| 6 | | |
| 7 | | |
| 7 | | |
| 7 | | |
| 8 | | |
| 9 | | |
| 10 | | |

My Notes

**9.** Calculate the sum of the deviations from the mean.

**10.** Explain the numerical value that you calculated in Item 9.

The next step is to divide by the standard deviation.

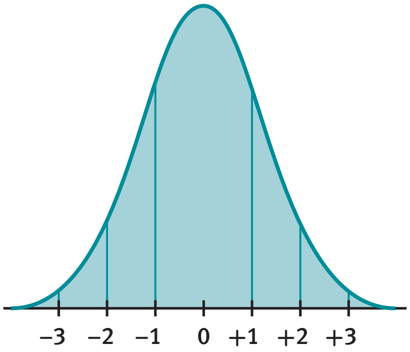
**11.** The standard deviation $s$ for Mr. Trent's class is 2.4. Use this to complete the last column of the table in Item 8. Round values to the nearest thousandth, if necessary.

### MATH TERMS

A ***z* score** is a standard score that indicates by how many standard deviations a data value is above or below the mean.

The numbers in the last column of the table represent the standardized scores for each data value. A standardized score is called a ***z* score:** $z = \frac{x - \overline{x}}{s}$. The set of $z$ scores is a data set with mean 0 and standard deviation 1. This represents a standardized version of the original data set.

**12.** Specific $z$ scores of $\pm1$ are indicated with vertical lines on the graph below. Draw vertical lines to indicate $z$ scores of $\pm2$ and $\pm3$.



The graph above shows a standardized normal distribution, which can be used to find probabilities. For example, the probability that a data value lies between –1 and +1 standard deviation from the mean is about 68%.

Use a normal distribution and a calculator to estimate the probability that a randomly selected student from Mr. Trent's class slept more than 7 hours. The steps for one type of graphing calculator are:

**Step 1.** Press 2nd VARS.

**Step 2.** Select normalcdf(.

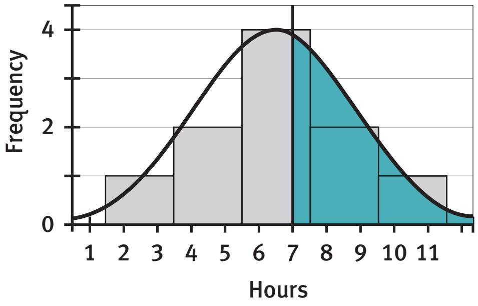
**Step 3.** Determine the lower boundary and enter that number, 7.

**Step 4.** Determine the upper boundary and enter that number, 10.

**Step 5.** Enter the mean, 6.5.

**Step 6.** Enter the standard deviation, 2.4.

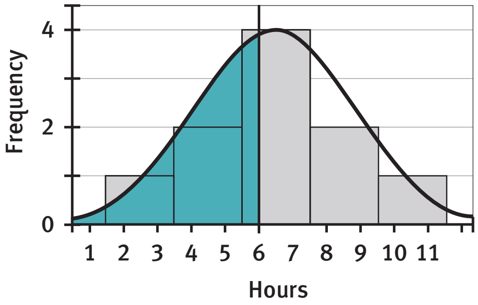
**Step 7.** Press Enter.



**13.** Using a calculator, find this probability.

**14.** Compare your answer to Item 13 with your answer to Item 6.

**15.** **Make sense of problems.** Use a normal distribution and a graphing calculator to estimate the probability that a student in Mr. Trent's class slept fewer than 6 hours. Does the answer that the calculator gives make sense when you look at the data? Explain.



**16.** Estimate the probability that a student in Mr. Trent's class slept fewer hours than you did in the last 24 hours.

My Notes

## Check Your Understanding

**17.** How many students from your class slept fewer hours than you did in the last 24 hours?

**18.** What percent of the students in your class slept between 6 and 8 hours in the last 24 hours? Explain how you would find this value and then calculate it.

**19.** You have discussed several examples of data sets that are normally distributed. Give an example of a real-world data set that would not be normally distributed and explain why.

**20.** The weights of 1.69-oz bags of M&Ms are normally distributed with a mean of 1.69 oz and a standard deviation of 0.05 oz. What is the probability that a bag selected at random weighs more than 1.76 oz?

**21.** There are yellow, blue, green, orange, red, and brown candies in every bag of M&Ms. The colors are normally distributed and, on average, 20% of the candies in a bag are blue, with a standard deviation of 5%. What is the probability that a bag of 27 M&Ms contains fewer than 5 blue candies?

**22.** Ian scored a 78 on last week's algebra test. The scores for the class were normally distributed with an average of 72 and a standard deviation of 5. What proportion of students scored higher than Ian?

## LESSON 37-3 PRACTICE

**Model with mathematics.** The times between eruptions of Old Faithful, a geyser at Yellowstone National Park, vary from 44 to 122 minutes. The average time between eruptions is 91 minutes. The table below lists the times between eruptions for January 1, 2011.

| Time in Minutes | |
|---|---|
| 85 | 85 |
| 85 | 92 |
| 100 | 88 |
| 85 | 90 |
| 99 | 100 |
| 91 | 101 |
| 86 | 96 |

**23.** Draw a dot plot and calculate the mean and standard deviation.

**24.** Determine the interval that represents 1 standard deviation on either side of the mean. Calculate the proportion of data values that lie in this interval, and show this on your dot plot.

**25.** Use a normal distribution to estimate the proportion of data values that lie between 85 and 92 minutes.

**26.** Compare your answers to Items 24 and 25.

**27.** Write a statement that explains how a normal distribution is related to the data on eruptions of Old Faithful.

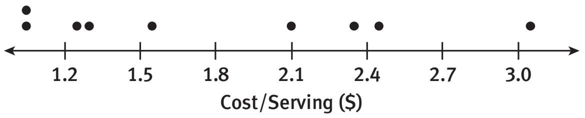
## ACTIVITY 37 PRACTICE

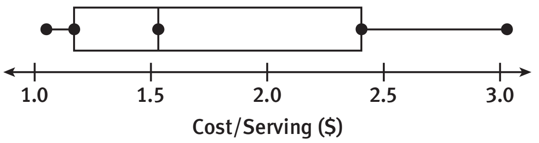
**Write your answers on notebook paper. Show your work.**

A restaurant specializing in Mexican food offers nine different choices of enchiladas. They vary in cost and in sodium content. Data for the nine options are given in the table below.

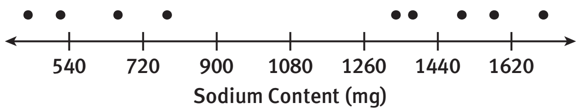
| Option | Cost/Serving (dollars) | Sodium Content (mg) |
|---|---|---|
| 1 | 3.03 | 780 |
| 2 | 1.07 | 1570 |
| 3 | 1.28 | 1500 |
| 4 | 1.53 | 1370 |
| 5 | 1.05 | 1700 |
| 6 | 1.27 | 1330 |
| 7 | 2.34 | 440 |
| 8 | 2.47 | 520 |
| 9 | 2.09 | 660 |

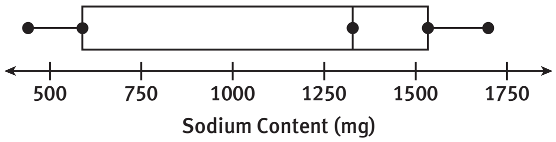
**1.** A dot plot and a box plot of the Cost/Serving amounts are shown below. What feature of the data set is apparent in the box plot but not particularly apparent in the dot plot?



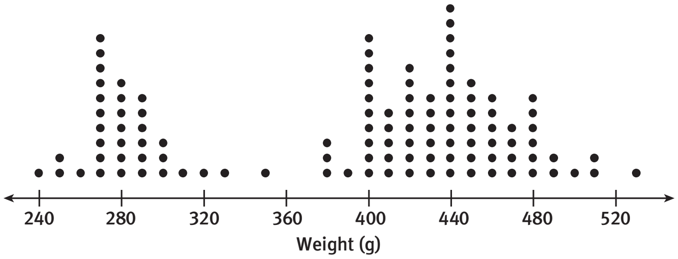


**2.** A dot plot and a box plot of the Sodium Content amounts are shown below. What feature of the data set is hidden in the box plot but apparent in the dot plot?





**3.** In a study of raptors in the western United States, 110 Cooper's hawks were trapped and their weights (in grams) recorded. A dot plot of these weights is shown below. What interesting feature do you notice about this data set? What do you think might be the reason for this interesting feature?



For winter sports enthusiasts, the thickness of ice is a significant safety issue. The Minnesota Department of Natural Resources recommends that ice thickness be at least 4 inches for walking or skating on the ice, and at least 5 inches for operating a snowmobile or all-terrain vehicle on the ice. Ice thicknesses (in inches) were measured at 10 randomly selected locations on the surface of a lake. The thicknesses were as follows:

5.8, 6.4, 6.9, 7.2, 5.1, 4.9, 4.3, 5.8, 7.0, 6.8

**4.** Construct a dot plot of the ice thicknesses.

**5.** On the basis of your dot plot, do you think it is safe to play hockey on this lake? Explain why or why not.

**6.** On the basis of your dot plot, do you think it is safe to operate a snowmobile on this lake? Explain why or why not.

**7.** Calculate the mean ice thickness for the locations in this sample.

**8.** Calculate the standard deviation of the ice thicknesses.

**9.** If the mean of the thicknesses were greater and the standard deviation were the same, would you be more worried or less worried about operating a snowmobile on the ice on this lake? Explain.

**10.** If the mean of the thicknesses were the same and the standard deviation were greater, would you be more worried or less worried about walking or skating on the ice on this lake? Explain.

Distributors of soft drinks are aware that end-aisle displays in stores are effective for increasing sales. A distributor is testing new designs of displays, where the image on the display is varied. Each image pictures one or two smiling people holding an open container of the soft drink. The distributor would like to know which images increase sales the most. The three different images are one man, one woman, and a pair of individuals (one man and one woman). Each image was used in 11 stores for 1 month, and the percent increases in sales compared to the same month of the previous year were recorded. Data from the 33 stores are shown in the table below.

**Percent Increase in Sales**

| Image: One Man | Image: One Woman | Image: Man and Woman |
|---|---|---|
| 4.79 | 5.71 | 8.18 |
| 5.71 | 6.29 | 9.14 |
| 5.74 | 7.44 | 9.70 |
| 5.54 | 6.03 | 9.25 |
| 4.43 | 5.54 | 7.40 |
| 6.42 | 5.90 | 9.25 |
| 6.07 | 5.23 | 8.42 |
| 3.97 | 7.96 | 8.12 |
| 5.84 | 4.75 | 9.07 |
| 5.55 | 4.68 | 7.84 |
| 6.76 | 5.90 | 8.09 |
| 5.62 | 5.71 | 9.18 |

**11.** For each of these three data sets, calculate the five-number summary and the upper and lower boundaries for outliers.

| | Man | Woman | Both |
|---|---|---|---|
| Minimum | | | |
| First quartile | | | |
| Median | | | |
| Third quartile | | | |
| Maximum | | | |
| Lower outlier boundary | | | |
| Upper outlier boundary | | | |

**12.** Use the information from the table in Item 11 to sketch modified box plots of these three data sets. Be sure to indicate any outliers.

**13.** In a few sentences, describe the similarities and differences among the three data sets.

**14.** Which image would you recommend that the distributor use and why?

A regional symphony orchestra needs money to repair their theater, which was seriously damaged by flooding. They have tested three different methods of asking for donations: mail, phone, and direct appeal at social gatherings. Each method was used with 11 potential donors, and the amounts donated for each method are shown below. Use the table for Items 15–17.

**Contributions ($)**

| Mail | Phone | Direct |
|---|---|---|
| 1000 | 1700 | 900 |
| 1500 | 1800 | 1000 |
| 1200 | 1900 | 1200 |
| 1800 | 1750 | 1500 |
| 1600 | 2000 | 1200 |
| 1100 | 1700 | 1550 |
| 1000 | 1800 | 1000 |
| 1250 | 1850 | 1100 |
| 1400 | 1500 | 1250 |
| 1300 | 900 | 1250 |
| 1400 | 1400 | 1350 |

**15.** Complete the table below.

**Data Summary Table ($)**

| | Mail | Phone | Direct |
|---|---|---|---|
| Minimum | | | |
| First quartile | 1100 | 1500 | 1000 |
| Median | 1300 | 1750 | 1200 |
| Third quartile | 1500 | 1850 | 1350 |
| Maximum | | | |
| Lower outlier boundary | | | |
| Upper outlier boundary | | | |

**16.** Construct modified box plots for the different methods.

**17.** Based on the data and your box plots, which method would you recommend and why?

**18.** Sometimes it is not clear whether a box plot is a modified box plot or a standard box plot. If you were looking at a box plot and outliers were not visible, what characteristics of the plot would lead you to believe it was standard rather than modified?

The following values represent the number of states visited by students in a class:

3, 12, 17, 2, 21, 14, 14, 8, 45, 29
Use these data for Items 19–22.

**19.** Find the interquartile range and any outliers for the data set.

**20.** If you found an outlier in Item 19, what does this number represent? Does it make sense that this number would be an outlier in this context? Explain your answer.

**21.** Create a modified box plot for the data.

**22.** Two new students joined the class, both of whom have visited only two states each. What effects, if any, does this have on the upper and lower boundaries for outliers?

## MATHEMATICAL PRACTICES
### Use Appropriate Tools Strategically

**23.** In Item 12 of Lesson 37-1, you were asked about advantages of using box plots and dot plots to describe and compare distributions of scores. Do you think the advantages you found would exist not only for these data, but for numerical data in general? Explain.

# Correlation

ACTIVITY 38

## What's the Relationship?

## Lesson 38-1 Scatter Plots

My Notes

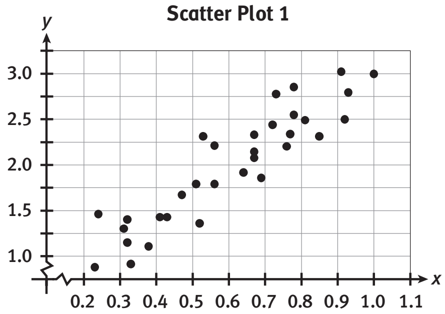
### Learning Targets:

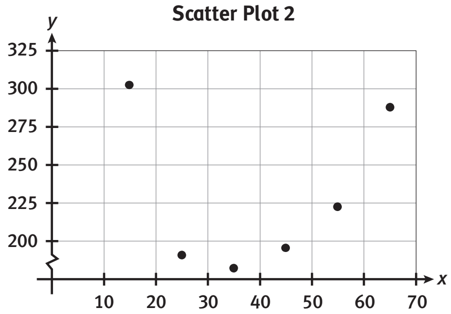
- Describe a linear relationship between two numerical variables in terms of direction and strength.
- Use the correlation coefficient to describe the strength and direction of a linear relationship between two numerical variables.

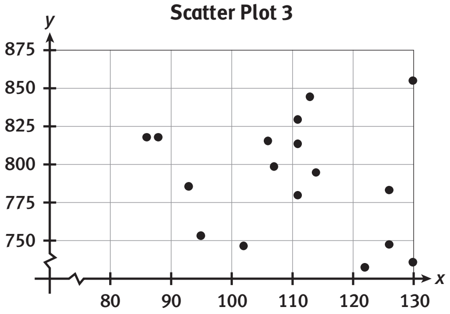
**SUGGESTED LEARNING STRATEGIES:** Graphic Organizer, Think-Pair-Share, Create Representations, Predict and Confirm, Quickwrite

**Scatter plots** are used to visualize the relationship between two numerical variables. When you look at a scatter plot, determine whether there appears to be a relationship (pattern) between the two variables.

For example, consider the following three scatter plots.


Scatter Plot 1


Scatter Plot 2


Scatter Plot 3

For Scatter Plot 1, there does appear to be a relationship between *x* and *y* because greater values of *x* tend to be paired with greater values of *y*. Notice that the pattern in the plot looks roughly linear, so you would say that there is a linear relationship between these two variables.

Two numerical variables are related if they tend to vary together in a predictable way.

**1.** For Scatter Plot 2 above, does there appear to be a relationship between *x* and *y*? If so, describe the pattern.

**2.** For Scatter Plot 3 above, does there appear to be a relationship between *x* and *y*? If so, describe the pattern.

### MATH TERMS

Two numerical variables are **correlated** if one variable tends to increase (or decrease) as the other variable increases.

### MATH TERMS

The **correlation coefficient** is a measure of the strength and direction of a linear relationship.

The correlation coefficient is denoted by *r*.

### MATH TIP

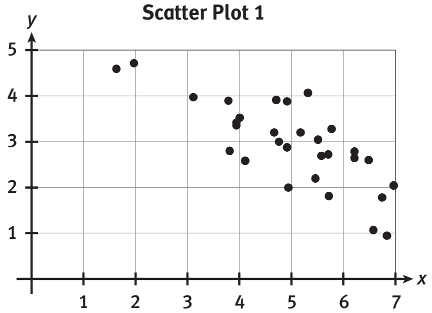
Variables are **positively related** if lesser values of one variable tend to occur with lesser values of the other variable.

Variables are **negatively related** if lesser values of one variable tend to occur with greater values of the other variable.

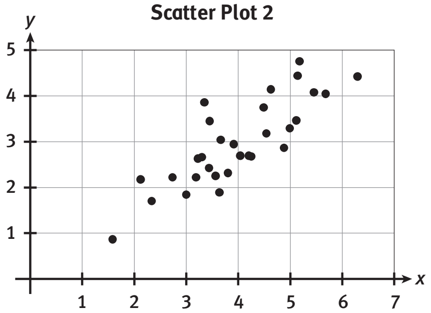
Just as the mean and standard deviation are used to describe center and variability in a data set, there is a summary statistic to describe the strength (how close the points are to a line) and direction (positive or negative) of a linear relationship. This statistic is called the **correlation coefficient** and is denoted by *r*.

**3.** Given below are seven scatter plots and seven verbal descriptions of relationships. Match each scatter plot with the appropriate description. (Each scatter plot goes with one and only one description.)

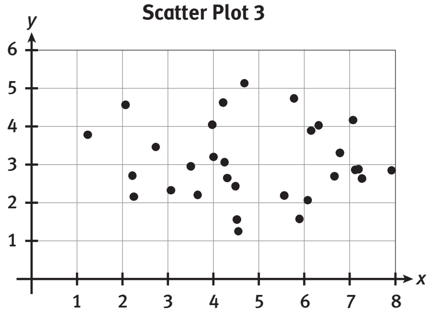
Scatter Plot 1



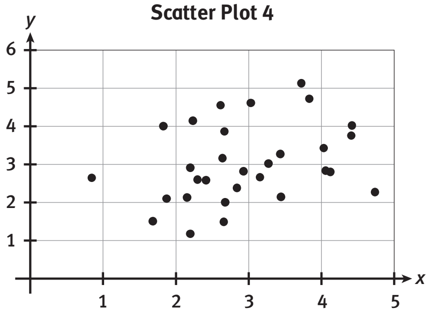
Scatter Plot 2



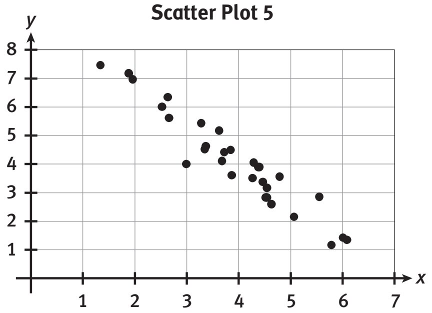
Scatter Plot 3



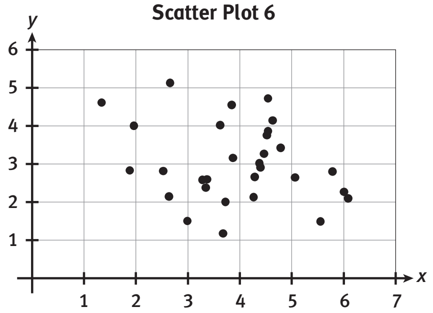
Scatter Plot 4



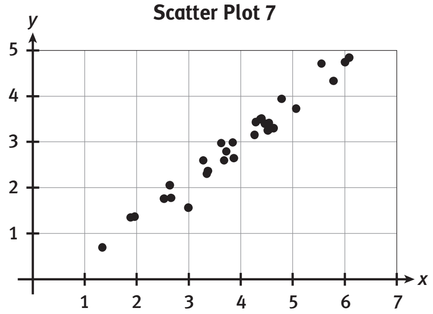
Scatter Plot 5



Scatter Plot 6



Scatter Plot 7



My Notes

**A.** Very strong positive linear relationship ($r = 0.981$) ________

**B.** Relatively strong positive linear relationship ($r = 0.828$) ________

**C.** Relatively weak positive linear relationship ($r = 0.310$) ________

**D.** Very slight or no linear relationship ($r = 0.043$) ________

**E.** Relatively weak negative linear relationship ($r = -0.238$) ________

**F.** Relatively strong negative linear relationship ($r = 0.772$) ________

**G.** Very strong negative linear relationship ($r = -0.95$) ________

**4.** What feature(s) of the scatter plots did you consider when deciding whether a relationship was positive or negative?

**5.** What feature(s) of the scatter plots did you consider when deciding whether a relationship was relatively weak, relatively strong, or very strong?

**6.** **Make sense of problems.** Examine the values of $r$ for each relationship in Item 3. How does the value of $r$ relate to the scatter plots? What makes $r$ increase or decrease?

Here is a summary of important characteristics of $r$:

- The value of $r$ quantifies the strength of a linear relationship.
- The sign of $r$ describes the direction of the relationship: positive or negative.
- $r$ ranges in value between $-1$ (perfect negative linear relationship) and $+1$ (perfect positive linear relationship).

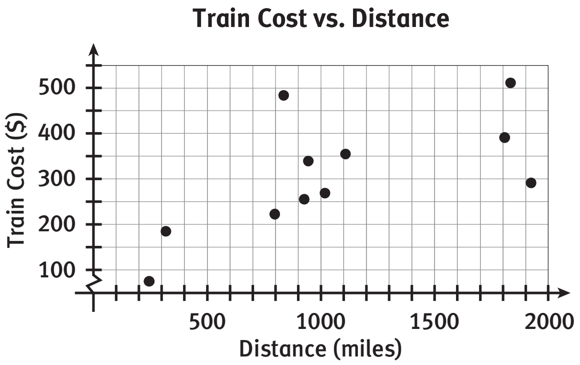
My Notes

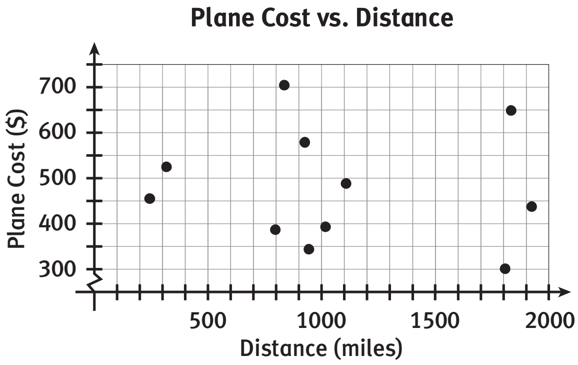
## Check Your Understanding

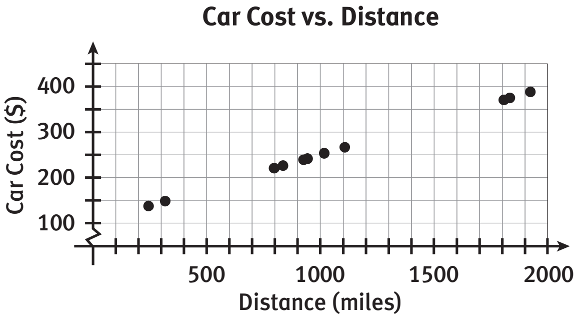
The following table displays costs to travel, round-trip, to various cities from Cedar Rapids, Iowa. The costs are calculated assuming a June 1 departure and a 3-day stay. Driving costs were calculated based on $0.20 per mile.

| | Travel Cost (dollars) | | |
|---|---|---|---|
| **Destination** | **Train** | **Plane** | **Car** |
| New York City | 268 | 391 | 204 |
| Chicago | 74 | 453 | 49 |
| Atlanta | 483 | 703 | 168 |
| Washington, D.C. | 254 | 577 | 186 |
| New Orleans | 338 | 342 | 189 |
| Denver | 221 | 384 | 160 |
| Albuquerque | 354 | 486 | 222 |
| Seattle | 510 | 647 | 367 |
| San Francisco | 290 | 435 | 385 |
| Los Angeles | 390 | 299 | 362 |
| Kansas City | 184 | 523 | 64 |

Scatter plots of cost versus distance for each of the three travel methods are shown below.

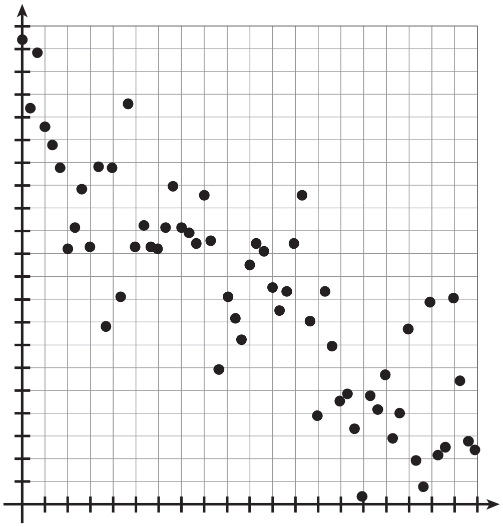






7. **Reason abstractly.** How would you describe the relationship between cost and distance for each method of transportation? Be sure to indicate whether you think the relationship is linear and to comment on the strength and direction of the relationship.
   **a.** Train
   **b.** Plane
   **c.** Car

## LESSON 38-1 PRACTICE



8. Describe the relationship shown in the scatter plot above.

9. **Reason abstractly.** In your own words, describe the similarities and differences between a scatter plot that shows a strong positive relationship and a scatter plot that shows a weak positive relationship.

10. What type of relationship would you expect to see between height and age? Explain your answer.

11. Describe two real-world quantities that would have a strong negative relationship.

12. Describe two real-world quantities that would have no correlation.

13. For positive linear relationships, as the value of *r* increases, is the linear relationship getting stronger or weaker?

## Learning Targets:

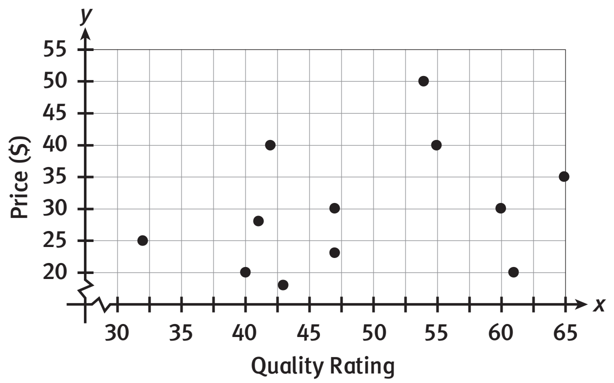
- Calculate correlation.
- Distinguish between correlation and causation.

**SUGGESTED LEARNING STRATEGIES:** Think-Pair-Share, Vocabulary Organizer, Quickwrite

The calculation of *r* gives additional information that helps to describe the data.

This data set shows price (in dollars) and quality ratings for 12 different brands of bike helmets. The quality rating is a number from 0 (worst) to 100 (best) that measures various factors such as how well the helmet absorbed the force of an impact, the strength and ventilation of the helmet, and its ease of use.

| Bicycle Helmets | Price (dollars) | 35 | 20 | 30 | 40 | 50 | 23 | 30 | 18 | 40 | 28 | 20 | 25 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Quality Rating | 65 | 61 | 60 | 55 | 54 | 47 | 47 | 43 | 42 | 41 | 40 | 32 |



**1. a.** How would you describe the relationship between price and quality rating?

**b.** Make a prediction of what you think the correlation coefficient might be.

**2.** Using a graphing calculator, enter the prices as one list and the quality ratings as another list. What is the value of the correlation coefficient for these two variables?

## Lesson 38-2
### Correlation Coefficient

My Notes

3. **Reason abstractly.** How would you interpret the value of the correlation coefficient in the context of this problem?

At this point you have used scatter plots to visually represent the relationship between two numerical variables and you have used a numerical measure to describe the strength and direction of a linear relationship. When a relationship is uncovered by statistics, the next task is to explain its meaning.

Sometimes the interpretation of a relationship may not be obvious. For example, across European countries there is a positive linear relationship between the number of storks and the number of newborn babies. Do storks bring babies? Are storks attracted by babies? Are both babies and storks brought by the tooth fairy? Do parents with newborns have warmer houses, and therefore their chimneys attract storks looking for warm places to nest?

Humans want to make sense of their world, and sometimes leap too quickly from seeing a correlation to inferring *causation* (a cause-and-effect relationship between two variables). This tendency should be resisted! There are many reasons why two variables might be related other than cause and effect.

Here are some common examples where a correlation should not be interpreted as a cause-and-effect relationship:

- The number of fire engines responding to a fire is positively correlated with the total damage. (Should fewer fire engines—perhaps 0—be sent to fires to reduce damage?)
- The number of people drowning at beaches is positively correlated with ice cream sales. (Is ice cream dangerous?)
- Shoe size is strongly correlated with reading ability. (Should parents start their children off with size 12?)
- The number of doctors per 1000 people is positively correlated with the rate of serious disease. (Are doctors spreading disease?)

4. **Make sense of problems.** For each of the correlations above, what do you think is the correct explanation for the correlation?

Be sure to use common sense when determining causation.

My Notes

## Check Your Understanding

**5.** For each of the following pairs of variables, indicate whether you would expect a positive correlation, a negative correlation, or a correlation close to 0. Explain your choice.

**a.** Weight of a car and gas mileage

**b.** Size and selling price of a house

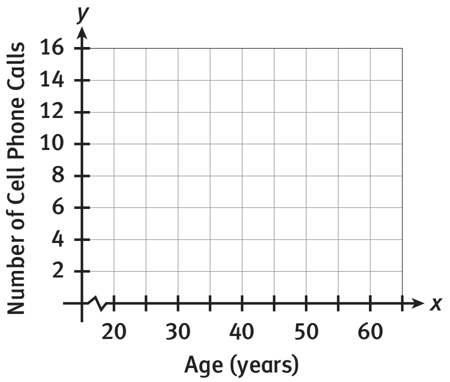
**c.** Height and weight

**d.** Height and number of siblings

The table below gives data on age and number of cell phone calls made in a typical day for each person in a random sample of 10 people. Use the table for Items 6–8.

| Age (years) | Number of Cell Phone Calls |
|---|---|
| 55 | 6 |
| 33 | 5 |
| 60 | 1 |
| 38 | 2 |
| 55 | 4 |
| 19 | 15 |
| 30 | 10 |
| 33 | 3 |
| 37 | 3 |
| 52 | 5 |

**6.** Sketch a scatter plot of these data using the grid below.



My Notes

**7.** Describe the direction and strength of the relationship between these two variables.

**8.** Calculate the value of the correlation coefficient. Do the sign and the magnitude of the correlation coefficient agree with your answer in Item 7? Explain.

**9.** Suppose that you shot an arrow into the air and kept track of how high it was every 1.0 second. If you made a scatter plot of the data (time, height), the resulting pattern of points would be in the shape of a parabola. Do you feel the correlation coefficient should be used to describe the strength of the relationship between time and height? Why or why not?

## LESSON 38-2 PRACTICE

**Model with mathematics.** Consider tablet computers with 9- to 10-inch screens.

**10.** For tablet computers, do you think there is a relationship between price and battery life? If so, do you think the relationship is positive or negative?

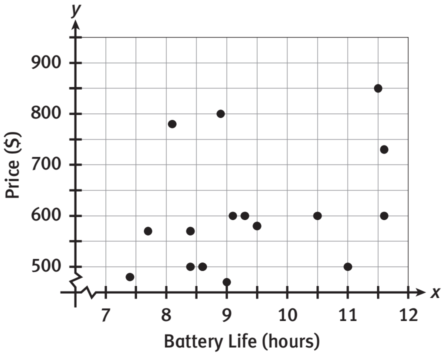
**11.** For tablet computers, do you think there is a relationship between price and weight? If so, do you think the relationship is positive or negative?

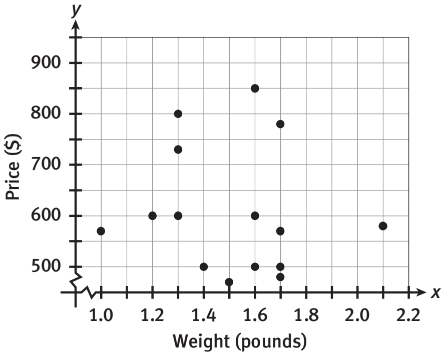
My Notes

Data for tablet computers with 9- to 10-inch screens are shown in the table and scatter plots below.

**9- to 10-inch Tablets**

| Price (dollars) | Battery Life (hours) | Weight (pounds) |
|---|---|---|
| 730 | 11.6 | 1.3 |
| 570 | 8.4 | 1.0 |
| 600 | 11.6 | 1.3 |
| 600 | 9.3 | 1.2 |
| 600 | 9.1 | 1.3 |
| 800 | 8.9 | 1.3 |
| 600 | 10.5 | 1.6 |
| 850 | 11.5 | 1.6 |
| 500 | 11.0 | 1.6 |
| 470 | 9.0 | 1.5 |
| 500 | 8.6 | 1.4 |
| 500 | 8.4 | 1.6 |
| 480 | 7.4 | 1.7 |
| 500 | 8.6 | 1.7 |
| 570 | 7.7 | 1.7 |
| 780 | 8.1 | 1.7 |
| 580 | 9.5 | 2.1 |





**12.** Calculate the correlation coefficient for Price and Battery Life.

**13.** Calculate the correlation coefficient for Price and Weight.

**14.** Are the values of the correlation coefficients consistent with your predictions in Items 10 and 11? Explain.

## ACTIVITY 38 PRACTICE

**Write your answers on notebook paper.**
**Show your work.**

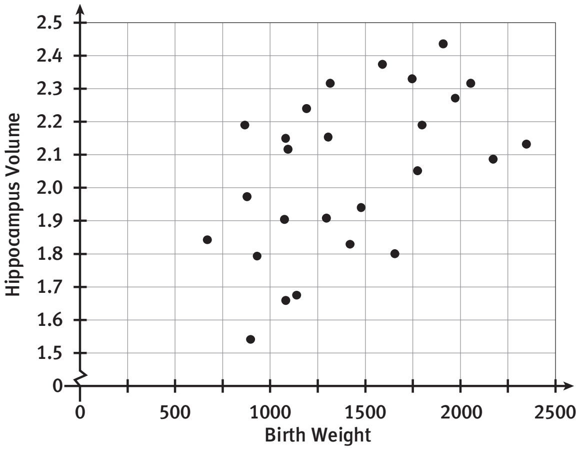
1. Describe as precisely as you can how the appearance of a scatter plot showing a positive linear relationship between two quantitative variables differs from the appearance of a scatter plot showing a negative relationship between two quantitative variables.
2. Describe as precisely as you can how the appearance of a scatter plot showing a strong linear relationship between two quantitative variables differs from the appearance of a scatter plot showing a weak linear relationship between two quantitative variables.

The basking shark is the second-largest fish (after the whale shark) swimming in the oceans today. In a study of these creatures, their length and average swimming speed were measured from a safe distance. The results are shown in the table below. Use the table for Items 3–5.

| Body Length (meters) | Average Speed (meters/sec) |
|---|---|
| 4.0 | 0.89 |
| 4.5 | 0.83 |
| 4.0 | 0.76 |
| 6.5 | 0.94 |
| 5.5 | 0.94 |

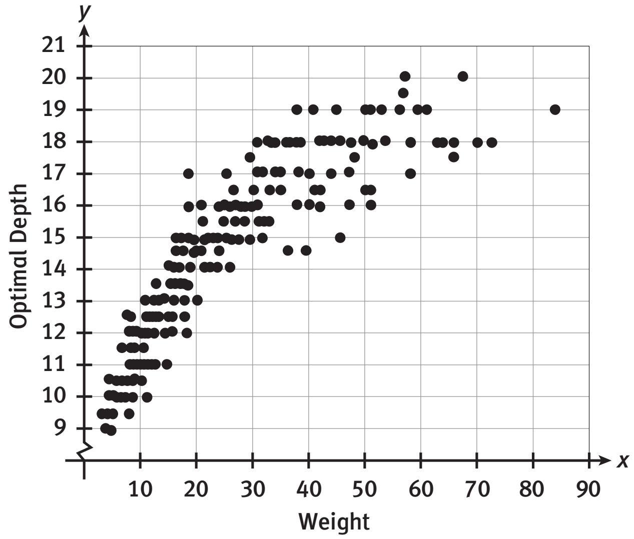
3. Sketch a scatter plot of these data.
4. Calculate the correlation coefficient for these data using your available technology.
5. How would you describe this relationship in terms of strength and direction? Support your description with specific references to the scatter plot and/or the correlation coefficient.

One danger of premature human birth is low birth weight. It is thought that low birth weight results in small hippocampus volume, which might be cause for concern because the hippocampus is important in later brain functioning. The scatter plot below displays data from a study of the relationship between hippocampus volume and birth weight in premature infants. The correlation coefficient for these data is $r = 0.51$.



6. Describe the strength and direction of this relationship.
7. Does this relationship appear to be reasonably described as linear? Explain.

When young children are prepared for surgery, a tracheal tube is inserted to allow the unconscious child to breathe. It is very important to get the correct insertion depth. Researchers investigated the relationship between best insertion depth and the weight of the child in a large sample of children, and a scatter plot of their data is shown. The correlation coefficient is $r = 0.878$.



**8.** Describe the strength and direction of this relationship.

## MATHEMATICAL PRACTICES

### Reason Abstractly and Quantitatively

**9.** Does this relationship appear to be reasonably described as linear? Why or why not?

# The Best-Fit Line

## Regressing Linearly

## Lesson 39-1 Line of Best Fit

ACTIVITY 39

### Learning Targets:

- Describe the linear relationship between two numerical variables using the best-fit line.
- Use the equation of the best-fit line to make predictions and compare the predictions to actual values.

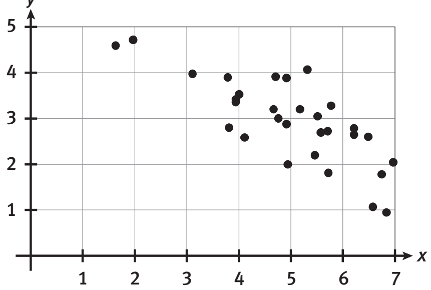
**SUGGESTED LEARNING STRATEGIES:** Look for a Pattern, Interactive Word Wall, Predict and Confirm, Graphic Organizer, Discussion Groups

In recent activities, you created scatter plots as a way to graphically summarize bivariate numerical data. In addition, you learned how to use the correlation coefficient as a numerical summary of the strength and direction of a linear relationship.

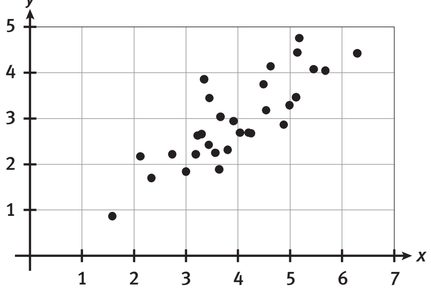
In this activity, you will see a way to summarize bivariate numerical data called the "best-fit line." You will use technology to determine the slope and *y*-intercept of the best-fit line for a data set.

1. The scatter plots below show linear relationships of different strengths and directions. For each scatter plot, use your judgment to draw a line that you feel best represents the linear relationship.

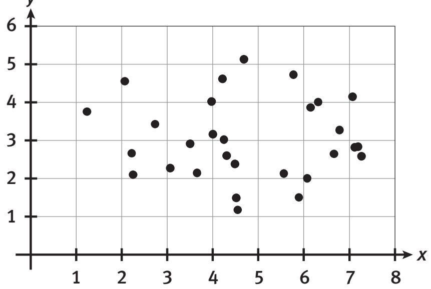
**Scatter Plot 1**



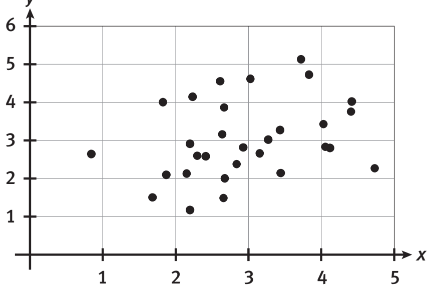
**Scatter Plot 2**



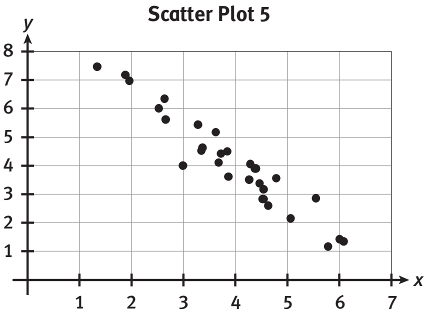
**Scatter Plot 3**

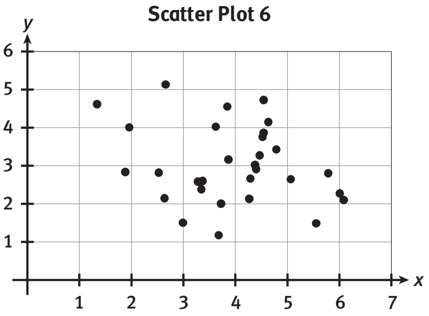


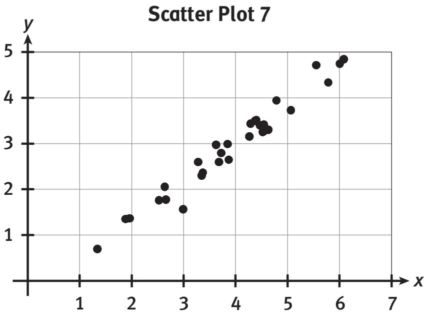
**Scatter Plot 4**



My Notes







2. Compare the lines you drew with the lines drawn by another student in your class. Did you draw identical lines? Were your lines more similar for scatter plots where the linear relationship was strong or where the linear relationship was weak?

Because informal assessments of what line might best describe a linear relationship don't always agree, we need to come to some agreement about what "best" means.

Before we look at how to define the best-fit line, let's first consider how the best-fit line might be used.

One reason for finding a best-fit line to describe the relationship between two variables is so that you can use the line to make predictions. For example, you might want to predict the age (in years) of a black bear from its weight (in pounds). This would be helpful to wildlife biologists, because it is a lot easier to weigh a bear than to ask a bear its age!

## Lesson 39-1
### Line of Best Fit

My Notes

Suppose you know that for adult black bears, the relationship between age and weight can be approximately described by the line

$$y = -3.69 + 0.115x$$

where $y$ = age in years and $x$ = weight in pounds. You can use this equation to predict the age of a bear that weighs 100 pounds:

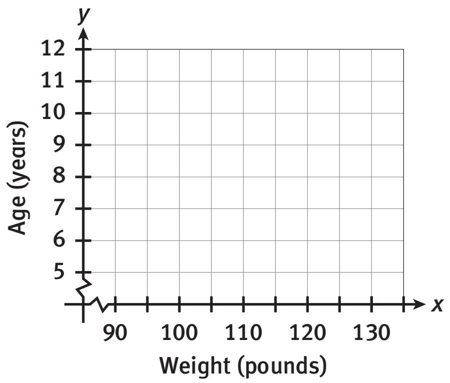
$$\text{predicted age} = -3.69 + 0.115(100) = -3.69 + 11.5 = 7.81 \text{ years}$$

**3.** Using the equation $y = -3.69 + 0.115x$, what is the predicted age of a bear that weighs 115 pounds?

The line $y = -3.69 + 0.115x$ is the best-fit line for the following data. These data are from a study in which nine black bears of known age were weighed.

| Bear | Weight ($x$) | Age ($y$) |
|---|---|---|
| 1 | 88.2 | 6.5 |
| 2 | 88.2 | 7.5 |
| 3 | 92.6 | 5.5 |
| 4 | 110.3 | 8.0 |
| 5 | 112.5 | 10.5 |
| 6 | 112.5 | 9.5 |
| 7 | 119.1 | 10.5 |
| 8 | 121.3 | 9.0 |
| 9 | 130.1 | 11.5 |

**4.** Construct a scatter plot for the bear data.

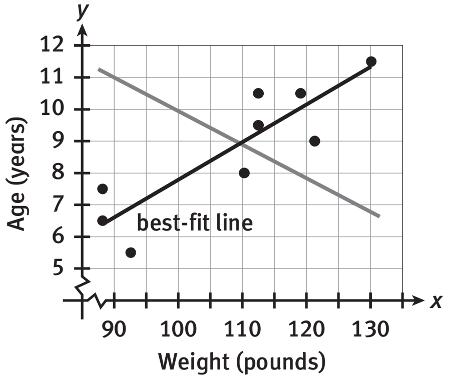


**5.** Add the best-fit line to your scatter plot.

(*Hint*: Find two points on the line by picking two $x$ values and using the equation of the best-fit line to find the corresponding predicted ages. Then plot these two ($x$, predicted age) pairs on the scatter plot and draw the line that goes through those two points.)

My Notes

Below is a scatter plot of the bear data and a line that is not the best-fit line.



**6.** Why is the best-fit line a better description of the relationship between age and weight than the line graphed above?

**7.** One bear in the data set (Bear 3) was 5.5 years old and weighed 92.6 pounds. If you used the best-fit line ($y = -3.69 + 0.115x$) to predict the age of this bear based on its weight, how far off would you be from the bear's actual age?

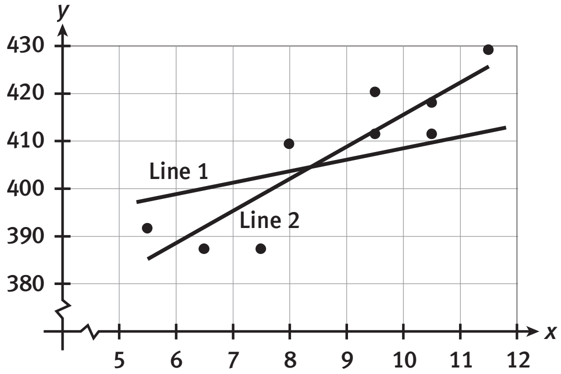
**8.** If you used the line graphed above to predict the age of this bear, do you think your prediction would be closer to or further from the bear's actual age? What feature(s) of the scatter plot shown above supports your answer?

**9.** **Attend to precision.** For the bear that was 5.5 years old and weighed 92.6 pounds, the best-fit line led to a predicted age that was greater than the bear's actual age. Will age predictions based on the best-fit line be greater than the actual age for *all* of the bears in the data set? If so, explain why. If not, give an example of a bear in the data set for which the predicted age is less than the bear's actual age.

My Notes

## Check Your Understanding

**10.** The scatter plot below shows two lines, Line 1 and Line 2. One of these lines is the best-fit line. Which one is it?



**11.** Suppose that for students taking a statistics class, the best-fit line for a data set where $y$ is a student's test score (out of 100 points) and $x$ is the number of hours spent studying for the test is $y = 43 + 12x$.

**a.** What is the predicted test score for a student who studied for one hour?

**b.** What is the predicted test score for a student who studied for three hours?

My Notes

## LESSON 39-1 PRACTICE

Mr. Trent examined some data on head height and a person's actual height and found that a person's height is about 7.5 times his or her head height. "Head height" refers to the distance from the top of the head to the bottom of the chin. Using the data he gathered, Mr. Trent found that the equation of the best-fit line is $y = 2.5 + 7.5x$, where $y$ represents height in inches and $x$ represents head height in inches. Use this equation for Items 12–14.

**12.** Xavier's head height is 8.3 inches. Predict Xavier's height.

**13.** Tamisha is 5 foot 3 inches tall. Predict her head height.

**14.** **Reason quantitatively.** Tori said that she is 64 inches tall and her head height is 8 inches. Is this possible? Explain.

## Lesson 39-2
### Residuals

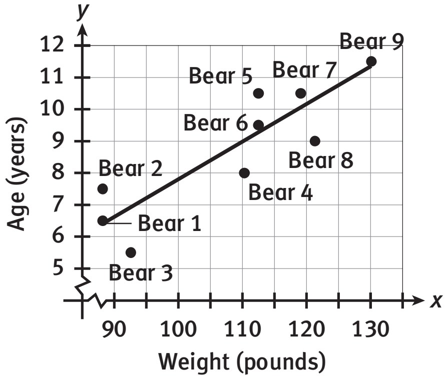
My Notes

**Learning Targets:**

- Use technology to determine the equation of the best-fit line.
- Describe the linear relationship between two numerical variables using the best-fit line.
- Use residuals to investigate whether a given line is an appropriate model of the relationship between numerical variables.

**SUGGESTED LEARNING STRATEGIES:** Predict and Confirm, Look for a Pattern

Below is a scatter plot of the bear data with the best-fit line and the points in the scatter plot labeled according to which bear the data point represents.



**1.** For which bears does the best-fit line predict an age that is less than the bear's actual age?

**2.** Look at the points in the scatter plot that correspond to the bears whose predicted ages are less than their actual ages. What do the points all have in common relative to the best-fit line?

## Lesson 39-2
### Residuals

For Bear 3, the actual age was 5.5 years and the predicted age from the best-fit line was 6.96 years. The difference between the actual age and the predicted age is

$$5.5 - 6.96 = -1.46$$

**3.** Look at the scatter plot and locate the point corresponding to Bear 3. What does 1.46 represent in terms of the scatter plot?

The difference between an actual $y$-value and a predicted $y$-value is called a **residual**. A residual is positive when the actual $y$-value is greater than the predicted $y$-value.

**MATH TERMS**

A **residual** is a difference between an actual $y$-value and a predicted $y$-value.

Residual = actual $y$ − predicted $y$

**4.** When is a residual negative?

**5.** For which of the bears is the residual positive?

**6.** Look at the scatter plot. Do data points that fall above the best-fit line have positive or negative residuals?

The table below shows the actual ages, predicted ages using the best-fit line, and residuals for the nine bears.

| Bear | Age (years) | Predicted Age (years) | Residual |
|---|---|---|---|
| 1 | 6.5 | 6.45 | 0.05 |
| 2 | 7.5 | 6.45 | 1.05 |
| 3 | 5.5 | 6.96 | −1.46 |
| 4 | 8.0 | 9.00 | −1.00 |
| 5 | 10.5 | 9.25 | 1.25 |
| 6 | 9.5 | 9.25 | 0.25 |
| 7 | 10.5 | 10.01 | 0.49 |
| 8 | 9.5 | 10.26 | −1.26 |
| 9 | 11.5 | 11.27 | 0.23 |

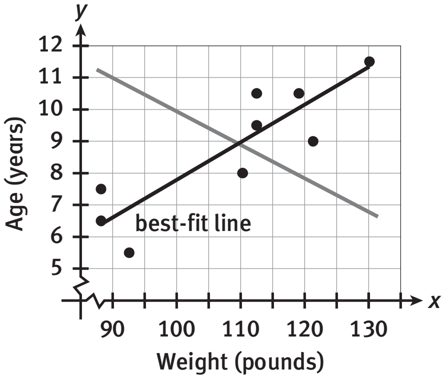
**7. Make sense of problems.** What is the sum of all nine residuals? Does this value surprise you? Explain why or why not.

*Note*: The sum of the residuals for the best-fit line is equal to zero. Here, because of rounding in the calculation of the slope and $y$-intercept of the best-fit line and rounding in calculating the predicted values, the sum of the residuals is not exactly 0.

## Lesson 39-2
### Residuals

My Notes

**8.** The scatter plot below shows the best-fit line and another line. If you ignore the sign of the residuals, which line has greater residuals overall? (*Hint*: Look at the distances of points to each of the two lines.)



A line is a good description of a bivariate data set if the residuals tend to be small overall. To measure the overall "goodness" of a line, you might think about adding all of the residuals. The problem with this is that some residuals are positive and some are negative, and so you can get a sum that is zero (or close to zero) even for lines that are not good descriptions of the data. So, instead of judging the "goodness" of a line by looking at the sum of the residuals, we look at the **sum of the squared residuals** (**SSR**, for short). The squared residuals are all positive, so positive and negative values don't offset one another.

**MATH TIP**

**SSR** stands for the sum of the squared residuals.

**9.** Look again at the scatter plot above that shows the bear data and the two different lines. Which line do you think has the lesser SSR? Explain your reasoning.

The **best-fit line** for a particular data set is the line that has the least sum of squared residuals (least SSR). In the scatter plot with the two lines, not only does the best-fit line have an SSR less than that of the other line shown, it has an SSR less than that of *any* other line.

**MATH TERMS**

The **best-fit line** is the line for which the sum of squared residuals (SSR) is less than that of any other line.

Calculating the equation of the best-fit line by hand is very time-consuming, especially if there are a lot of values in the data set. Because of this, you will use a graphing calculator or computer software to do the calculations.

**10.** Enter the bear data and use technology to verify that the equation of the best-fit line is $y = -3.69 + 0.115x$.

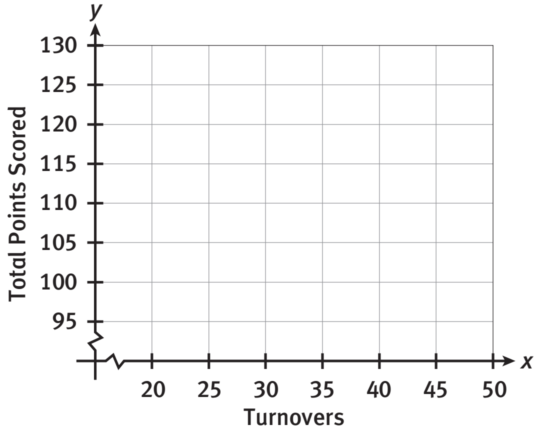
My Notes

## Check Your Understanding

The men's basketball coach at Grinnell College employs a style of basketball known as "system ball." The idea behind system ball is that forcing turnovers on defense leads to more shots, especially 3-point shots, on offense, and thus a higher point total. Data on the number of turnovers committed by the opposing team and the total points scored by Grinnell for a sample of seven games are given below.

| Turnovers ($x$) | Total Points Scored ($y$) |
|---|---|
| 36 | 115 |
| 45 | 126 |
| 26 | 103 |
| 18 | 106 |
| 25 | 117 |
| 31 | 128 |
| 22 | 96 |

**11.** Construct a scatter plot for this data set.



**12.** Based on the scatter plot, how would you describe the relationship between $x$ and $y$?

**13.** Use technology to find the equation of the best-fit line.

**14.** Use the best-fit line to predict the total points scored in a game with 30 turnovers.

## LESSON 39-2 PRACTICE

My Notes

The table below shows the historical minimum wage (in dollars per hour) for the State of New York. Use the table for Items 15–18.

| Year ($x$) | Wage ($y$) |
|---|---|
| 1962 | 1.15 |
| 1968 | 1.60 |
| 1974 | 2.00 |
| 1978 | 2.65 |
| 1981 | 3.35 |
| 1990 | 3.80 |
| 1991 | 4.25 |
| 2000 | 5.15 |
| 2005 | 6.00 |
| 2012 | 7.25 |

**15.** Construct a scatter plot of the data.

**16.** Does there appear to be a relationship between the year and the minimum wage? If so, describe the relationship.

**17.** **Use appropriate tools strategically.** Find the equation for the best-fit line and the correlation coefficient.

**18.** **Construct viable arguments.** Do the equation you found and the correlation coefficient support your answer in Item 16? Explain.

My Notes

## Learning Targets:

- Interpret the slope of the best-fit line in the context of the data.
- Distinguish between scatter plots that show a linear relationship and those where the relationship is not linear.

**SUGGESTED LEARNING STRATEGIES:** Quickwrite, Look for a Pattern, Visualization, Create Representations, Guess and Check

Once you have determined the equation of the best-fit line, it is possible to interpret the slope of the line. For the bear data, the slope of the best-fit line is 0.115. This means that for each additional pound of weight, the predicted age increases by 0.115 years.

### MATH TIP

The slope of the best-fit line can be interpreted as the change in the value of the predicted *y* variable that is associated with a change of 1 in the value of the *x* variable.

**1.** The best-fit line for a data set where $y$ is the time to complete a task (in seconds) and $x$ is the room temperature (in degrees Fahrenheit) is $y = 128 + 2x$. The slope of this line is 2. Interpret this value in the context of this problem.

**2.** What does it mean when the slope of the best-fit line is negative?

## Lesson 39-3
### Interpreting the Slope and Intercept of the Best-Fit Line

My Notes

**3.** The best-fit line for a data set where $y$ is the fuel efficiency of a car (in miles per gallon) and $x$ is the weight of the car (in pounds) is $y = 40 - 0.005x$. The slope of this line is $-0.005$. Interpret this value in the context of this problem.

It is sometimes also possible to interpret the $y$-intercept of the best-fit line, but this is not always a sensible thing to do. The $y$-intercept of the line is the point whose $x$-coordinate is 0. But it doesn't make sense to predict the age of a bear whose weight is 0 or the fuel efficiency of a car that weighs 0 pounds. So, in most cases, you won't want to interpret the $y$-intercept.

**When is it NOT okay to use the best-fit line to make predictions?**

There are three situations when it doesn't make sense to use the best-fit line to make predictions.

You should not use the best-fit line

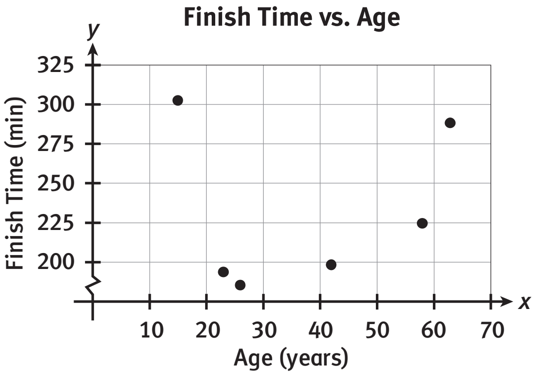
- to predict a value that is far outside the range of values in the data set used to find the best-fit line.
- when the linear relationship between $x$ and $y$ is very weak, which means that the residuals are very large overall.
- when the relationship between $x$ and $y$ is not linear and would be better described by a curve.

**4. Make sense of problems.** Why do you think it is not a good idea to predict a value that is far outside the range of the data used to find the best-fit line? For example, why would it not be a good idea to predict the fuel efficiency of a 500-pound car if the data set had only cars that weighed between 2000 and 3500 pounds?

## Lesson 39-3
### Interpreting the Slope and Intercept of the Best-Fit Line

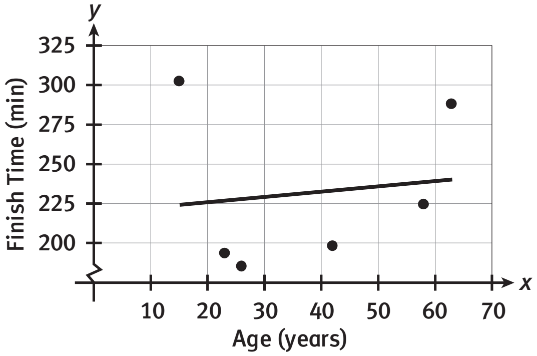
**5.** Even though the best-fit line has an SSR that is less than the SSR of any other line, the residuals might still be great. That is, the points in the data set might still tend to fall far from the line. If this is the case, do you think predictions based on the line will tend to be close to actual *y*-values? Explain your reasoning.

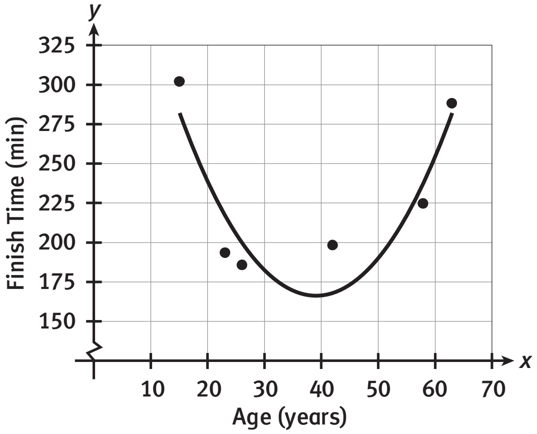
Sometimes there is a relationship between two numerical variables, but the relationship is not linear. For example, consider the scatter plot below. This scatter plot was constructed using data on *y*, the time to finish a marathon (in minutes), and *x*, the age (in years), for six women.



This scatter plot shows a nonlinear relationship between finish time and age.

6. Below are scatter plots that show the best-fit line and the best-fit quadratic curve. To predict finish time based on age, would you recommend using the best-fit line or the best-fit quadratic curve? Explain why you made this choice.





My Notes

**7.** What do you think it means to say a quadratic curve is the best-fit quadratic curve?

**8.** The equation of the best-fit quadratic curve for the marathon data is

$$\text{finish time} = 473.3 - 15.80(\text{age}) + 0.2030(\text{age})^2$$

What is the predicted finish time for a 30-year-old woman?

**9.** Would you recommend using the best-fit quadratic curve to predict the finish time for a woman who is 80 years old? Explain why or why not.

## Check Your Understanding

**10.** Explain why you should not use the best-fit line for the bear data to predict the age of a 40-pound bear.

**11.** Suppose a bear weighs 110.5 pounds. Would you recommend using the best-fit line to predict the age of this bear? Why or why not?

**12.** The Chestnut High School tennis team recorded the amount of time that each player's last tennis match lasted as well as the number of calories the player burned during the match. Dakota plays number 1 singles and her match lasted 3 hours; she burned 1900 calories. Briana is number 2 singles; she beat her opponent with no problem and her match lasted only 45 minutes. She burned 475 calories. Maria beat her opponent in 1.5 hours and she burned 950 calories. Determine the equation of the best-fit line and interpret the slope in the context of the problem.

## LESSON 39-3 PRACTICE

Use the following information for Items 13–16.

At a recent football game you noticed that students tend to be near the same height as their parent of the same gender. You surveyed several students and their parents and recorded the data below.

| Student's Height (in.) | Parent's Height (in.) |
|---|---|
| 61 | 56 |
| 62 | 57 |
| 63 | 62 |
| 65 | 66 |
| 65 | 60 |
| 66 | 66 |
| 67 | 68 |
| 68 | 63 |
| 68 | 69 |
| 70 | 71 |
| 72 | 68 |
| 71 | 73 |
| 73 | 67 |
| 74 | 71 |

**13.** **Model with mathematics.** Draw a scatter plot and describe the relationship between the students' heights and their parents' heights.

**14.** Using a graphing calculator, determine the equation of the best-fit line and calculate the correlation coefficient.

**15.** Use the best-fit line to predict the height of your classmate Tyler's father. Tyler is 70 inches tall.

**16.** **Attend to precision.** Tyler's father came to pick him up and you asked his height. He is 6 feet tall. What is his residual and where does the data point lie in relation to the best-fit line?

My Notes

## Lesson 39-4
## Plotting Residuals

### Learning Targets:

- Create a residual plot given a set of data and the equation of the best-fit line.
- Use residuals to investigate whether a line is an appropriate description of the relationship between numerical variables.

**SUGGESTED LEARNING STRATEGIES:** Create Representations, Look for a Pattern, Quickwrite

In the marathon data scatter plot from the previous lesson, it is obvious that the relationship between finish time and age is not linear. But sometimes, the nonlinearity of the relationship between two variables isn't always this obvious. One way to decide whether a curve describes a relationship better than a line is to look at a residual plot.

A **residual plot** is a scatter plot of the ($x$, residual) pairs.

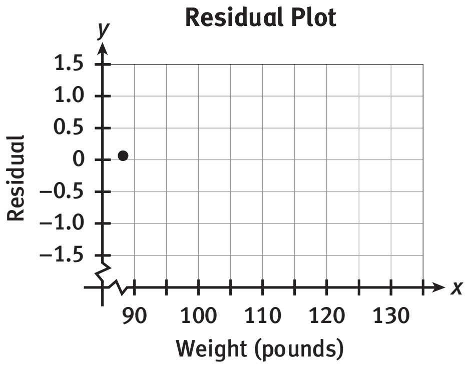
### MATH TERMS

A **residual plot** is a scatter plot of the ($x$, residual) pairs.

To see how to make a residual plot, let's return to the bear data. For the bear data set, the data and the residuals for the best-fit line are shown in the table below.

| Bear | Weight (pounds) | Age (years) | Predicted Age (years) | Residual |
|---|---|---|---|---|
| 1 | 88.2 | 6.5 | 6.45 | 0.05 |
| 2 | 88.2 | 7.5 | 6.45 | 1.05 |
| 3 | 92.6 | 5.5 | 6.96 | −1.46 |
| 4 | 110.3 | 8.0 | 9.00 | −1.00 |
| 5 | 112.5 | 10.5 | 9.25 | 1.25 |
| 6 | 112.5 | 9.5 | 9.25 | 0.25 |
| 7 | 119.1 | 10.5 | 10.01 | 0.49 |
| 8 | 121.3 | 9.5 | 10.26 | −1.26 |
| 9 | 130.1 | 11.5 | 11.27 | 0.23 |

The first ($x$, residual) pair is (88.2, 0.05). This point has been graphed below.



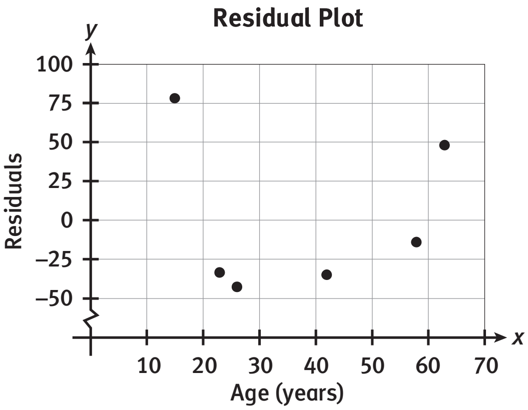
1. Complete the residual plot by adding the other eight ($x$, residual) points to the plot.

## Lesson 39-4
### Plotting Residuals

My Notes

Notice that there is no pattern in the residual plot for the bear data best-fit line. The points appear to be scattered at random in this plot.

Now look at a residual plot for the best-fit line for the marathon data set.



**MATH TIP**

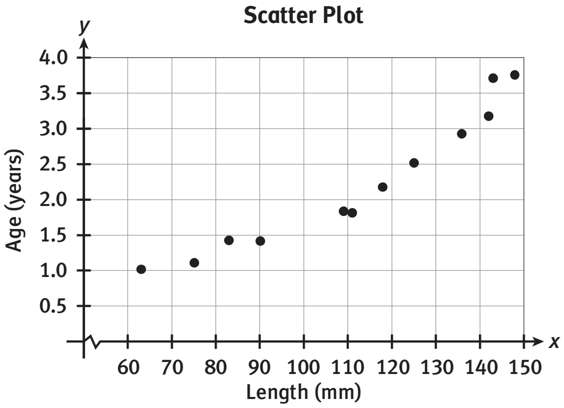
A strong pattern in the residual plot for the best-fit line indicates that a line is not the best way to describe the relationship.

Here there is a very strong curved pattern. It is this pattern in the residual plot that confirms that a line is **not** the best way to describe the relationship between finish time and age.

Now let's look at another example where biologists were interested in predicting the ages of lobsters. The data below are the shell lengths and ages for 12 lobsters.

| Shell Length (mm) | Age (years) |
|---|---|
| 63 | 1.00 |
| 83 | 1.42 |
| 109 | 1.82 |
| 111 | 1.80 |
| 118 | 2.17 |
| 143 | 3.70 |
| 90 | 1.40 |
| 125 | 2.51 |
| 136 | 2.92 |
| 75 | 1.10 |
| 142 | 3.17 |
| 148 | 3.75 |

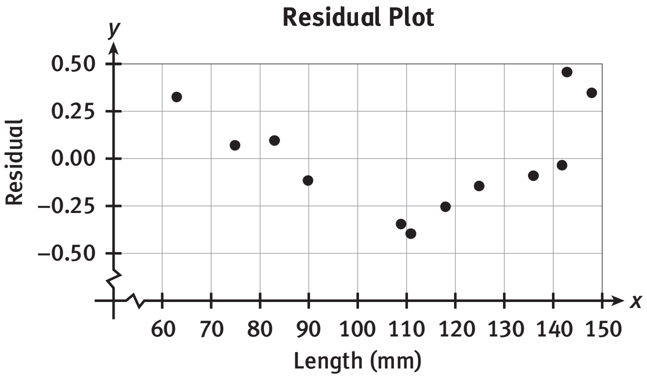
A scatter plot of the data is shown below.



**2.** The equation of the best-fit line for this data set is $y = -1.41 + 0.0325x$, where $y$ = age (in years) and $x$ = shell length (in mm). Use this equation to complete the following table. Round values to the nearest hundredth if necessary.

| Shell Length (mm) | Age (years) | Predicted Age (years) | Residual |
|---|---|---|---|
| 63 | 1.00 | 0.64 | 0.36 |
| 83 | 1.42 | | 0.13 |
| 109 | 1.82 | | −0.31 |
| 111 | 1.80 | 2.20 | −0.40 |
| 118 | 2.17 | 2.43 | |
| 143 | 3.70 | 3.24 | |
| 90 | 1.40 | 1.52 | −0.12 |
| 125 | 2.51 | 2.65 | −0.14 |
| 136 | 2.92 | 3.01 | −0.09 |
| 75 | 1.10 | | |
| 142 | 3.17 | | |
| 148 | 3.75 | | |

A residual plot (scatter plot of the $(x, \text{residual})$ pairs) is shown here:



## Lesson 39-4
### Plotting Residuals

My Notes

**3.** Describe the pattern in the residual plot.

**4.** **Critique the reasoning of others.** The biologists who collected these data decided to use an exponential curve to describe the relationship between age and shell length. Do you think this was a reasonable choice? Explain why or why not.

### Check Your Understanding

No tortilla chip lover likes soggy chips, so chip makers want to know what makes them crispy. The data below are from an experiment to see how the moisture content of tortilla chips is related to the frying time.

Here, $x$ = frying time (in seconds) and $y$ = moisture content (in percent).

| Frying Time ($x$) | Moisture Content ($y$) |
|---|---|
| 5 | 16.3 |
| 10 | 9.7 |
| 15 | 8.1 |
| 20 | 4.2 |
| 25 | 3.4 |
| 30 | 2.9 |
| 45 | 1.9 |
| 60 | 1.3 |

**5.** Construct a scatter plot for this data set.

**6.** Find the equation of the best-fit line.

My Notes

**7.** Compute the predicted values and the residuals.

| *x* | *y* | Predicted *y* | Residual |
|---|---|---|---|
| 5 | 16.3 | | |
| 10 | 9.7 | | |
| 15 | 8.1 | | |
| 20 | 4.2 | | |
| 25 | 3.4 | | |
| 30 | 2.9 | | |
| 45 | 1.9 | | |
| 60 | 1.3 | | |

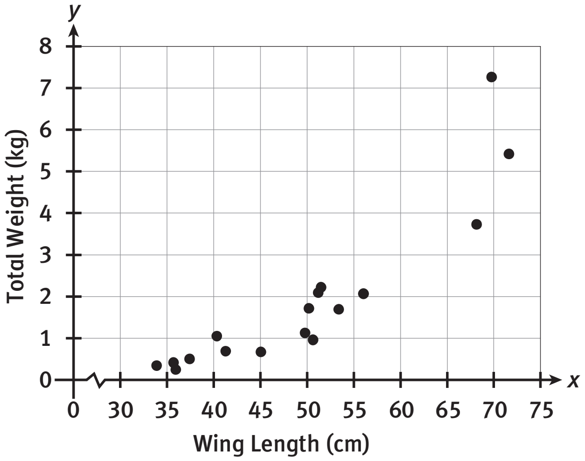
**8.** Construct a residual plot.

**9.** **Construct viable arguments.** Based on the scatter plot and the residual plot, would you recommend using the best-fit line to describe the relationship between *x* and *y*? Explain.

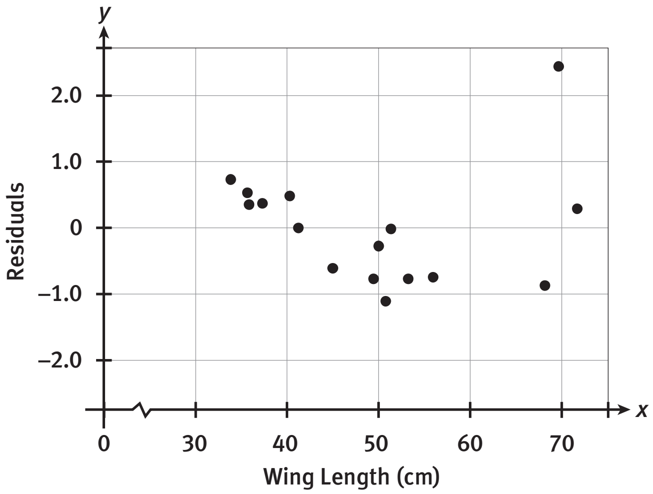
## LESSON 39-4 PRACTICE

**10.** Why is it important to look at the scatter plot and the residual plot before deciding whether it is appropriate to describe the relationship between two numerical variables using the best-fit line?

**11.** A recent study of birds of prey resulted in data on $x$ = wing length and $y$ = total weight for 16 different species. The scatter plot for these data is shown below. The correlation coefficient is $r = 0.897$. Describe the relationship between these variables in terms of strength, direction, and shape.



**12.** The residual plot for the data in Item 11 is shown below. Does the residual plot support your description of the relationship between these two variables? Explain, referring to specific characteristics of the residual plot.



## ACTIVITY 39 PRACTICE

**Write your answers on notebook paper. Show your work.**

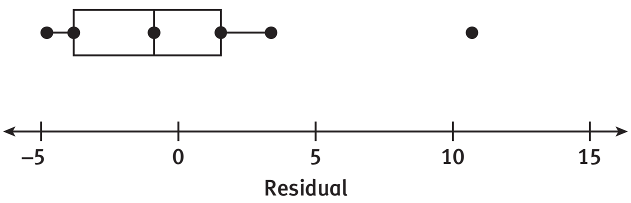
A veterinarian is studying the relationship between the weight of one-year-old golden retrievers in pounds ($y$) and the amount of dog food the dog is fed each day in pounds ($x$). A random sample of 10 one-year-old golden retrievers yielded the data in the table below. Use the table for Items 1–8.

| Dog | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| $x$ | 0.5 | 1.0 | 0.6 | 1.0 | 1.3 | 1.5 | 0.5 | 0.7 | 0.8 | 0.6 |
| $y$ | 42 | 67 | 47 | 55 | 62 | 71 | 36 | 42 | 50 | 43 |

1. Construct a scatter plot of the dog food data.

The equation of the best-fit line is $y = 24.6 + 31.7x$; the correlation coefficient is $r = 0.92$.

2. Interpret the value of the slope of the best-fit line.
3. Does it make sense to interpret the meaning of the $y$-intercept of the best-fit line?
4. One dog in this data set is fed 0.8 pound of food per day. If you used the best-fit line to predict the weight of this dog, how far off would you be from the dog's actual weight?
5. A box plot of the residuals is shown below. One of the residuals is an outlier. To which dog does this residual belong?



6. Suppose that you graphed the best-fit line on your scatter plot in Item 1. Looking at the scatter plot with the best-fit line, how would you know whether a point had a positive or a negative residual?
7. Suppose that the point with the greatest residual was the result of an error in data recording, and that the actual residual for that point is 2.0. Would the correlation increase or decrease? Explain your reasoning in a few sentences.
8. Below are the 10 residuals for the best-fit line. Construct a residual plot. Are there any patterns in the residual plot that indicate the relationship between $x$ and $y$ is not linear?

| Dog | Residual |
|---|---|
| 1 | 1.55 |
| 2 | 10.7 |
| 3 | 3.38 |
| 4 | −1.3 |
| 5 | −3.81 |
| 6 | −1.15 |
| 7 | −4.45 |
| 8 | −4.79 |
| 9 | 0.04 |
| 10 | −0.62 |

Data on the end of semester exam scores ($y$) and the number of hours spent studying ($x$) for 34 students in an algebra class were used to find the equation of the best-fit line. A scatter plot of these data showed a strong linear pattern. The equation of the best-fit line was $y = 42 + 11x$. Use this information for Items 9–11.

9. Interpret the value of the slope of the best-fit line.
10. Interpret the meaning of the $y$-intercept of the best-fit line.

## MATHEMATICAL PRACTICES
### Construct Viable Arguments and Critique the Reasoning of Others

11. Study times for these 34 students ranged from 0 to 6 hours. Explain why it is not reasonable to use the best-fit line to predict the test score of a student who studied for 10 hours.

# Bivariate Data

ACTIVITY 40

## Categorically Speaking

## Lesson 40-1 Bivariate Categorical Data

My Notes

### Learning Targets:

- Summarize bivariate categorical data in a two-way frequency table.
- Interpret frequencies and relative frequencies in two-way tables.

**SUGGESTED LEARNING STRATEGIES:** Summarizing, Paraphrasing, Create Representations

In previous activities, you analyzed bivariate numerical data. You measured the strength of association between two numerical variables and learned how to predict the value of one variable given the value of another, using the best-fit line. In this activity, you will work with bivariate categorical data.

The first example you will consider involves a famous data set from a tragic historical event: survival data from the luxury ship SS *Titanic*. The *Titanic* set sail on her maiden voyage on April 10, 1912. On a moonless night, April 12, she struck an iceberg and sank, and many people died. Hundreds of books and scholarly studies have tried to answer the question of how a ship so well constructed could have come to such a sad end, and government inquiries resulted in recommendations for future ship construction and safety regulations. This accident has been the topic of many popular movies, plays, and televisions specials.

Table 1 is taken from the United States Senate report dated May 28, 1912: Investigation into Loss of S. S. "Titanic". The format of the table has been preserved for historical accuracy.

**Table 1: Original Senate Data**

| | On board. | | | Saved. | | | Lost. | | | Percent saved. |
|---|---|---|---|---|---|---|---|---|---|---|
| | Women and children. | Men. | Total. | Women and children. | Men. | Total. | Women and children. | Men. | Total. | |
| Passengers | | | | | | | | | | |
| First class . . . . . | 156 | 173 | 329 | 145 | 54 | 199 | 11 | 119 | 130 | 60 |
| Second class . . . | 128 | 157 | 285 | 104 | 15 | 119 | 24 | 142 | 166 | 42 |
| Third class . . . . | 224 | 486 | 710 | 105 | 69 | 174 | 119 | 417 | 536 | 25 |
| Total passengers | 508 | 816 | 1,324 | 354 | 138 | 492 | 154 | 678 | 832 | . . . |
| Crew . . . . . . . . . | 23 | 876 | 899 | 20 | 194 | 214 | 3 | 682 | 685 | 24 |
| Total . . . . . . . . | 531 | 1,692 | 2,223 | 374 | 332 | 706 | 157 | 1,360 | 1,517 | 32 |

There are quite a few categorical variables summarized in this table: Person Type (men, women/children); Ticket Class (first, second, third); Role (passengers, crew); and Fate (saved, lost). You will use numbers from the Senate report table to form your own tables.

My Notes

1. Use the Senate report table to answer the following questions:
   a. How many first-class passengers were there?

   b. How many of the first-class passengers were men?

   c. How many male first-class passengers were saved?

2. **Reason quantitatively.** Compare the number of male second-class passengers saved with the number of male third-class passengers saved.

3. Was the percentage of second-class passengers saved greater than or less than the percentage of third-class passengers saved?

4. Based on your answers to Items 2 and 3, does it seem like the second-class or the third-class passengers were at greater risk of death?

To analyze these data we can consider two variables at a time. Bivariate categorical data are often summarized and arranged in a **two-way frequency table.** A two-way frequency table has rows and columns corresponding to the different possible values of the two categorical variables. For example, suppose you were interested in the variables Ticket Class and Fate. There are three values for Ticket Class and two values for Fate. The data can be summarized using a table with three rows and two columns. You can label the rows with the values for Ticket Class and the columns with the values for Fate.

MATH TERMS

A **two-way frequency table** summarizes the distribution of values for bivariate categorical data.

5. Use the information from the Senate report table to complete the two-way table Ticket Class vs. Fate.

**Ticket Class vs. Fate**

| | Saved | Lost |
|---|---|---|
| First class | 199 | |
| Second class | | |
| Third class | | |

Row and column totals are usually also included in a two-way table, as shown below. These totals are called *marginal* totals. The marginal totals describe the univariate distribution for each of the variables. For example, there were a total of 329 passengers with first-class tickets, and a total of 492 passengers were saved. The total number of passengers is shown in the bottom right column, the *grand total*.

**Ticket Class vs. Fate Frequencies**

*Cell counts*

| | Saved | Lost | Total |
|---|---|---|---|
| **First class** | 199 | 130 | 329 |
| **Second class** | 119 | 166 | 285 |
| **Third class** | 174 | 536 | 710 |
| **Total** | 492 | 832 | 1324 |

*Marginal totals* *Grand total*

It is common to convert these frequencies into *relative frequencies* by dividing by the grand total. These values are expressed in decimal form. As an example, consider the second-class passengers who were lost. The table in Item 5 shows that there were 119 second-class passengers who were saved. The grand total was 1324. Dividing 119 by 1324 gives 0.090 as the relative frequency.

**6.** Calculate the remaining relative frequencies and complete the two-way table below.

**Ticket Class vs. Fate**

| | Saved | Lost |
|---|---|---|
| **First class** | | |
| **Second class** | 0.090 | |
| **Third class** | | |

My Notes

## MATH TIP

*Frequency* refers to the number of times a particular value occurs in a data set.

*Relative frequency* is the proportion of the time that a particular value occurs in a data set.

*Marginal total* is the sum of frequencies or relative frequencies in a row or column of a two-way table.

My Notes

## Check Your Understanding

In recent years, there has been a great deal of controversy over the use of Native American team names and mascots for sports teams in the United States. *Sports Illustrated* magazine commissioned a survey of Native Americans living on reservations, Native Americans living off reservations, and non-Native-American sports fans to determine their feelings. One question asked about the "tomahawk chop" chant used in home games of the Atlanta Braves baseball team. Results are shown in the frequency table below.

**7.** Complete the frequency table below.

**Person vs. Attitude Toward "Tomahawk Chop"**
**Frequency Table**

| | Non-NA Fans | NA on Res | NA off Res | Total |
|---|---|---|---|---|
| **Like it** | 208 | 24 | 44 | |
| **Don't care** | 379 | | 66 | 545 |
| **Is objectionable** | 156 | 85 | 24 | 265 |
| **Total** | 743 | 209 | | |

**8.** Use the frequencies from the table above to complete the table below.

**Person vs. Attitude Toward "Tomahawk Chop"**
**Relative Frequency Table**

| | Non-NA Fans | NA on Res | NA off Res | Total |
|---|---|---|---|---|
| **Like it** | 0.192 | 0.022 | | 0.254 |
| **Don't care** | | 0.092 | 0.061 | 0.502 |
| **Is objectionable** | 0.144 | 0.078 | 0.022 | |
| **Total** | 0.684 | 0.192 | 0.123 | |

**9.** What is the relative frequency of non-Native-American fans who find the "tomahawk chop" objectionable?

**10.** What is the marginal frequency of Native Americans living on a reservation?

**11.** What is the marginal relative frequency of those who find the "tomahawk chop" objectionable?

My Notes

## LESSON 40-1 PRACTICE

**12.** Complete the following table on United States usage of multimedia devices in minutes per day.

**United States Usage of Multimedia in Minutes Per Day**

| Year | Web Browsing | Mobile Applications | Television | Total |
|---|---|---|---|---|
| 2010 | 70 | 66 | 162 | |
| 2011 | 72 | 94 | 168 | |
| 2012 | 70 | 127 | 168 | |
| Total | | | | |

**13.** Give the frequency of mobile application usage in 2012.

**14.** Determine the relative frequency of total time spent Web Browsing in 2010 for all usage of multimedia.

**15.** **Critique the reasoning of others.** Jayson states that people use multimedia devices of some type for 30% of the day. Do you agree with this statement? Why or why not?

## Lesson 40-2
### Presenting Relative Frequency Data Graphically

My Notes

### Learning Targets:

- Interpret frequencies and relative frequencies in two-way tables.
- Recognize and describe patterns of association in two-way tables.

**SUGGESTED LEARNING STRATEGIES:** Think-Pair-Share, Create Representations, Look for a Pattern

For most people, a visual presentation of data leads to faster understanding of the relationships between pairs of variables. A **segmented bar graph** is an effective way to present bivariate categorical data so that these relationships can be easily seen.

**MATH TERMS**

A **segmented bar graph** summarizes categorical data.

The total data set is represented by a bar, and the different possible categories are represented by sections of the bar. The area of the section for a particular category is proportional to the relative frequency of that category.

To see how a segmented bar graph is constructed, consider data from Major League Baseball.

Major League players are categorized as relief pitchers (RP), starting pitchers (SP), catchers (C), infielders (In), and outfielders (Out). The frequency table below summarizes data on position and team for the 75 players on three different teams (Cubs, Reds, and Pirates).

**Frequencies**
**Positions for Three Different Baseball Teams**

| | RP | SP | C | In | Out | Total |
|---|---|---|---|---|---|---|
| **Cubs** | 6 | 6 | 2 | 7 | 4 | 25 |
| **Reds** | 7 | 5 | 2 | 6 | 5 | 25 |
| **Pirates** | 7 | 4 | 2 | 8 | 4 | 25 |
| **Total** | 20 | 15 | 6 | 21 | 13 | 75 |

A segmented bar graph can be constructed by first calculating percentages within a row of data. Because we are interested in the percentages in each of the position categories for the different teams, we treat each individual row as a "whole." For each row, compute the percentages by first dividing each number in that row by the corresponding row total and then multiplying by 100.

**MATH TERMS**

**Row percentages** are calculated by dividing a number by the corresponding row total and then multiplying by 100.

The **row percentages** for the positions on the Cubs team are shown in the table below. Notice that the percentages in the Cubs row add to 100% because we are considering each row separately.

1. Use the frequencies in the table above to complete the following table.

**Row Percentages**
**Positions for Three Different Baseball Teams**

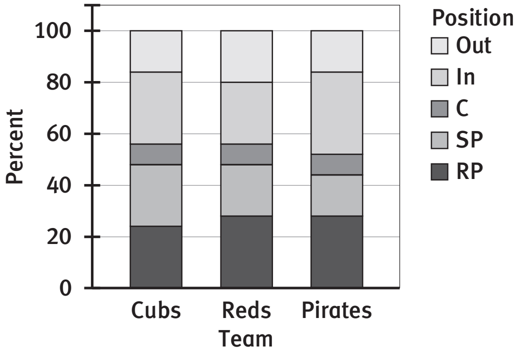
| | RP | SP | C | In | Out | Total |
|---|---|---|---|---|---|---|
| **Cubs** | 24% | 24% | 8% | 28% | 16% | 100% |
| **Reds** | | | | | | 100% |
| **Pirates** | | | | | | 100% |

## Lesson 40-2

### Presenting Relative Frequency Data Graphically

The segmented bar graph below shows the percentages of different positions (Outfield, Infield, etc.) for each of the three teams.

**Positions for Three Different Baseball Teams**



2. **Reason quantitatively.** Does it appear from the segmented bar graph that the distributions of positions are very similar for the three teams, or are they noticeably different? How are the distributions similar or different?

Two categorical variables are said to be ***associated*** if knowing the value of one of the variables gives you information about the value of the other variable. With categorical data we will ask whether the distribution of values for one variable is similar to or different from the distribution of values for the other variable. A segmented bar graph presents these distributions of values graphically. For the baseball example, the distributions of the position categories (distribution across the columns in each row) are very similar for all three teams (rows). We would say that the categorical variables of position and team are **not** associated. That is, knowing the percentage of pitchers on a team does not provide information about which team it is.

**ACADEMIC VOCABULARY**

To ***associate*** means to connect or to link.

The table below is a two-way frequency table for the variables Person Type and Fate on the *Titanic.*

**Person Type vs. Fate**

| | Saved | Lost | Total |
|---|---|---|---|
| **Woman/Child** | 354 | 154 | 508 |
| **Man** | 138 | 678 | 816 |
| **Total** | 492 | 832 | 1324 |

## Lesson 40-2
### Presenting Relative Frequency Data Graphically

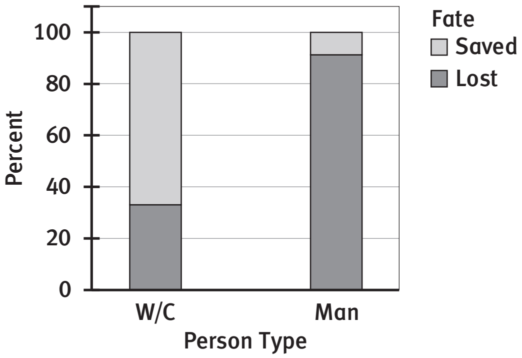
**3.** Calculate the percentages needed to construct a segmented bar graph by completing the table below.

**Person Type vs. Fate**

| | Saved | Lost | Total |
|---|---|---|---|
| **Woman/Child** | | | 100% |
| **Man** | | | 100% |

A segmented bar graph for these variables is shown below.

**Fate for Different Persons**



**4.** About what percent of the area of the bar for women and children (W/C) is ░? Does this correspond to the percentage of women and children who were saved?

**5.** About what percentage of the area of the bar for men is ▓? Does this correspond to the percentage of men who were lost?

Compare this segmented bar graph with the segmented bar graph for the baseball players. The segmented bars representing baseball positions were very similar for the three teams, indicating that the position and team are not associated. The segmented bars above are quite different in terms of the percentages of Lost and Saved. There is an association between Fate and Person Type because knowing the person type (for example, women and children) does provide useful information about fate.

## Lesson 40-2

### Presenting Relative Frequency Data Graphically

My Notes

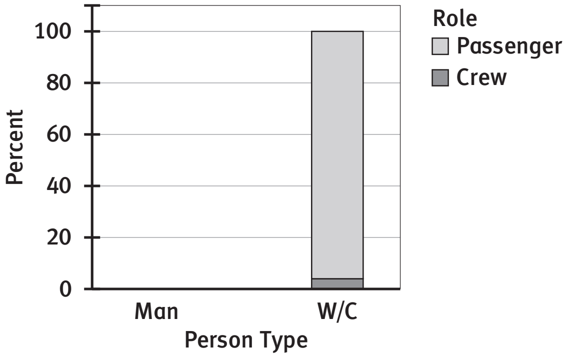
The table below is a two-way frequency table for the two variables Role and Person Type for the *Titanic* data.

| | Passenger | Crew | Total |
|---|---|---|---|
| Woman/Child | 508 | 23 | 531 |
| Man | 816 | 876 | 1692 |
| Total | 1324 | 899 | 2223 |

To construct a segmented bar graph, one variable is placed on the horizontal axis, and the vertical axis is labeled and scaled with percentages from 0% to 100%.

The two-way frequency table above shows that 508 of the 531 women and children (96%) were passengers and 23 of 531 (4%) were crew members. The vertical bar for women and children is "segmented" to reflect this distribution of percentages.

**Partial Segmented Bar Graph of Role vs. Person Type**



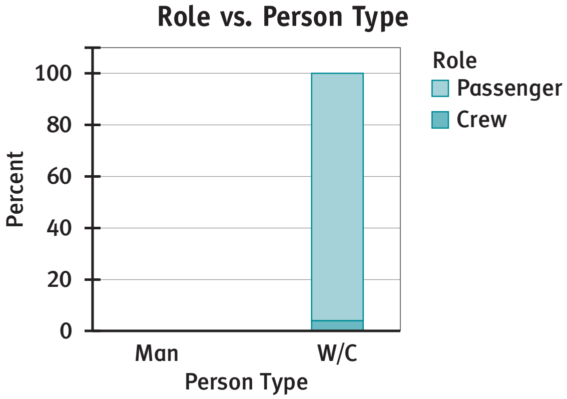
**6.** Complete the table below, rounding percentages to the nearest whole number.

**Role vs. Person Type**

| | Passenger | Crew | Total |
|---|---|---|---|
| Woman/Child | 96% | 4% | 100% |
| Man | | | 100% |

My Notes

**7.** Add the missing bar to the segmented bar graph below.



**8.** **Make sense of problems.** Comment on the association, if any, between these variables.

## Check Your Understanding

Kidney stones are solid lumps of crystals that separate from urine and build up on the inner surface of the kidney. If left untreated they can lead to kidney failure. Is there an association between sleep position and the location of kidney stones? Researchers asked patients with kidney stones to identify their preferred sleep positions. The table below classifies the responses by sleep position and kidney stone location.

**Frequencies of Kidney Stone Locations and Preferred Sleep Positions**

| | Left Kidney | Right Kidney | Total |
|---|---|---|---|
| **Right Side Down** | 31 | 13 | 44 |
| **No Preference** | 8 | 9 | 17 |
| **Left Side Down** | 9 | 40 | 49 |
| **Total** | 48 | 62 | 110 |

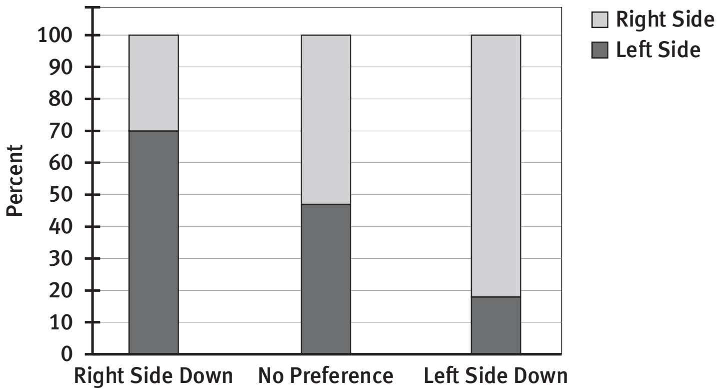
**9.** From the data summarized in the table above, find the row percentages that would be used to construct a segmented bar graph. Round the percentages to the nearest whole number and use them to complete the following table.

**Kidney Stone Locations and Preferred Sleep Positions**

| | Left Side | Right Side | Total |
|---|---|---|---|
| Right Side Down | 70% | | 100% |
| No Preference | | | 100% |
| Left Side Down | | 82% | 100% |

A segmented bar graph for the kidney stone data is shown below.

**Kidney Stone Locations and Preferred Sleep Positions**



**10.** **Make sense of problems.** Does there appear to be an association between sleep position and kidney stone location? What feature(s) of the segmented bar graph support your answer?

**11.** How is sleep position related to the location of kidney stones?

My Notes

## LESSON 40-2 PRACTICE

Does Vitamin C help prevent colds? In a study of 279 French skiers, 140 of them were given a placebo (a sham treatment with no active ingredients) and 139 of them were given Vitamin C. They were followed for one week and whether or not they caught a cold was recorded. The data from this study are shown below.

**Treatment vs. Cold**
**Frequency Table**

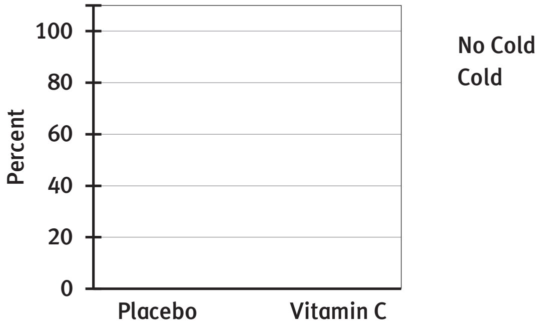
| | Cold | No Cold | Total |
|---|---|---|---|
| **Placebo** | 31 | 109 | 140 |
| **Vitamin C** | 17 | 122 | 139 |
| **Total** | 48 | 231 | 279 |

**12.** Find the row percentages that would be used to construct a segmented bar graph. Round the percentages to the nearest whole number and use them to complete the table below.

**Treatment vs. Cold**
**Percentages Across Rows**

| | Cold | No Cold | Total |
|---|---|---|---|
| **Placebo** | | | 100% |
| **Vitamin C** | | | 100% |

**13.** Sketch a segmented bar graph using the axes shown below.



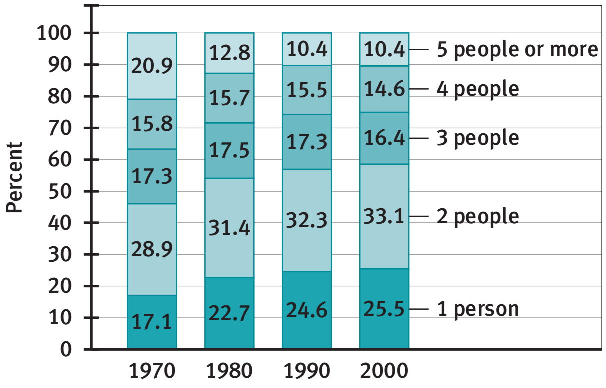
**14. Make use of structure.** It appears that there is an association between these two variables. Does this association suggest that Vitamin C might help prevent colds? What features of the data and the segmented bar graphs support your answer?

## ACTIVITY 40 PRACTICE

**Write your answers on notebook paper.
Show your work.**

As part of the United States Census, data are collected on the number of persons in each household. The census data for four decades are summarized below.

**Households by Size: Selected Years, 1970 to 2000**



Source: U.S. Census Bureau, Current Population Survey, March Supplements: 1970 to 2000.

1. Which size household increased the most on a percentage basis between 1970 and 2000?
2. Which size household decreased the most on a percentage basis between 1970 and 2000?

Myopia (nearsightedness) is a condition in which a person's eye is slightly longer than it should be, resulting in blurry images of objects far away. Hyperopia (farsightedness) is a condition in which a person's eye is slightly shorter than it should be, resulting in blurry images of near objects. There is some evidence that nighttime light exposure during sleep before the age of two years may be associated with myopia. Data from a survey of parents of children aged 2 to 16 seen at an eye clinic are summarized in the table below.

**Vision Condition vs. Light History Frequencies**

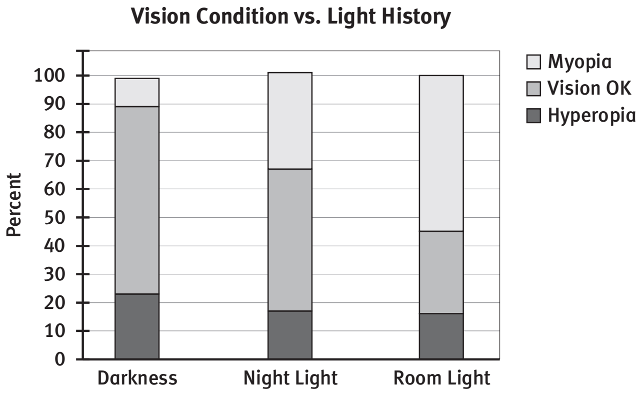
| | Darkness | Night Light | Room Light | Total |
|---|---|---|---|---|
| Hyperopia | 23 | 17 | 16 | 56 |
| Vision OK | 66 | 50 | 29 | 145 |
| Myopia | 10 | 34 | 55 | 99 |
| Total | 99 | 101 | 100 | 300 |

3. Use the information given in the table above to complete the table of relative frequencies below.

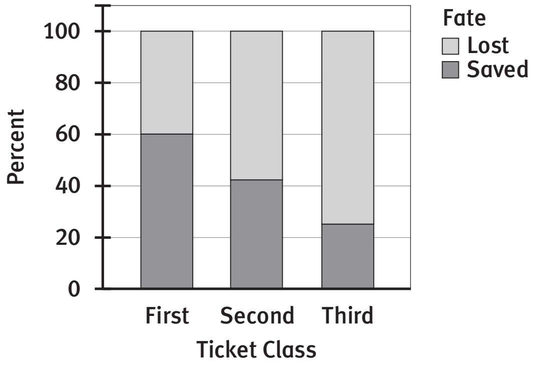
**Vision Condition vs. Light History Relative Frequencies**

| | Darkness | Night Light | Room Light | Total |
|---|---|---|---|---|
| Hyperopia | 0.077 | 0.057 | 0.053 | |
| Vision OK | 0.220 | | 0.097 | 0.483 |
| Myopia | | 0.113 | 0.183 | 0.330 |
| Total | 0.330 | 0.337 | | |

**4.** A segmented bar graph of Vision Condition vs. Light History is shown below. In a few sentences, describe the association between Vision Condition and Light History. (You may assume that room light has more light than a night light.)

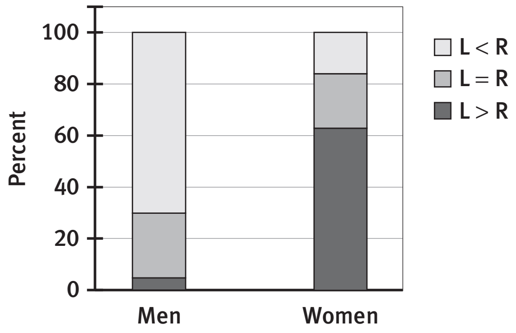


**5.** The segmented bar graph below shows the relationship between Ticket Class and Fate for those on the *Titanic*. Comment on the association between these two variables.



## MATHEMATICAL PRACTICES
### Make Sense of Problems and Persevere in Solving Them

**6.** In a study of right-handed men and women, data were gathered on gender and foot asymmetry. An individual was classified as having a left foot more than half a shoe size larger than the right foot ($L > R$), having a left foot more than half a shoe size smaller than the right foot ($L < R$) or having the same shoe size for both feet ($L = R$, which includes cases where both feet were within one half shoe size of each other). A segmented bar graph is shown below. Comment on the association between gender and foot asymmetry.


Percent
Men
Women
L < R
L = R
L > R