2

Gotta Keep It Correlatin'

Correlation

Warm Up

Describe a possible flaw in the reasoning for each situation.

1. If I wash my hands regularly, I will not get sick.

2. If I practice my guitar every day, I will be a rock star.

3. If I wear my favorite football jersey to support the team, they will win the game.

4. If I am a good driver, I will not have an accident.

Learning Goals

- Determine the correlation coefficient using technology.
- Interpret the correlation coefficient for a set of data.
- Understand the difference between r and r^2.
- Understand the difference between correlation and causation.
- Understand necessary conditions.
- Understand sufficient conditions.
- Choose a level of accuracy appropriate when reporting quantities.

Key Terms

- correlation
- correlation coefficient
- coefficient of determination
- causation
- necessary condition
- sufficient condition
- common response
- confounding variable

You have learned how to write a line of best fit using the Least Squares Method. How do you know if that line actually produces valid, useable results? Is there a way to measure the strength of the relationship between the variables?

Associate, Formulate, Correlate!

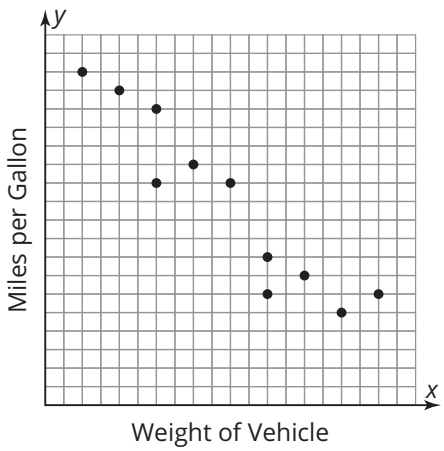
Consider each relationship shown.

1. Describe any associations between the independent and dependent variables, and then draw a line of best fit, if possible.

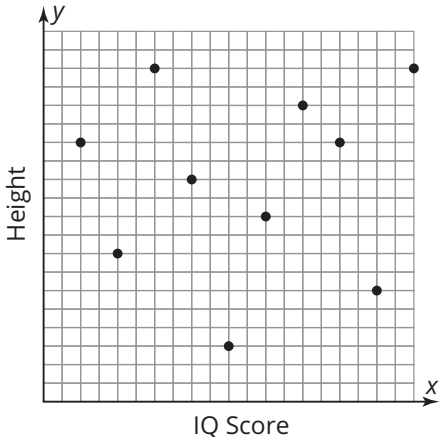
Remember:

Data comparing two variables can show a positive association, negative association, or no association.

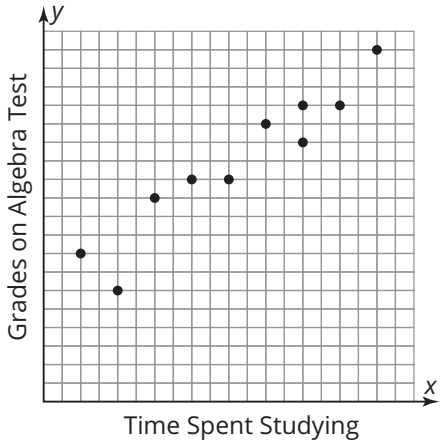
a.



b.



c.



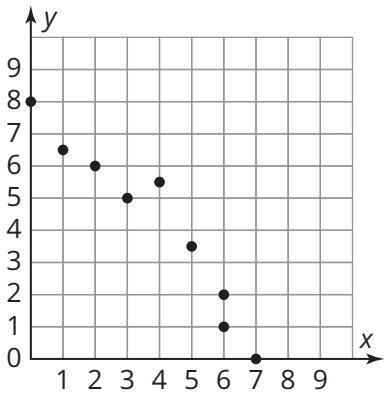
ACTIVITY
2.1 The Correlation Coefficient

A measure of how well a regression fits a set of data is called **correlation**. The **correlation coefficient** is a value between –1 and 1, which indicates how close the data are to the graph of the regression equation. The closer the correlation coefficient is to 1 or –1, the stronger the relationship is between the two variables. The variable *r* is used to represent the correlation coefficient.

The correlation coefficient falls between –1 and 0 if the data show a negative association or between 0 and 1 if the data show a positive association.

1. Determine whether the points in each scatter plot have a positive correlation, a negative correlation, or no correlation. Four possible *r*-values are given. Circle the *r*-value you think is most appropriate. Explain your reasoning for each.

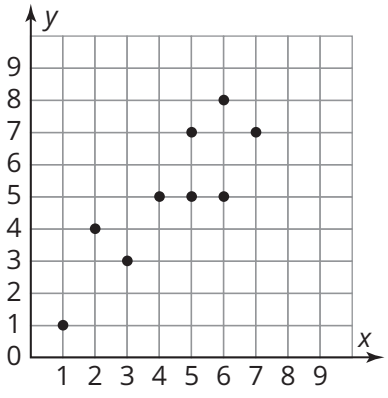
a.



- $r = 0.9$
- $r = -0.9$
- $r = 0.09$
- $r = -0.09$

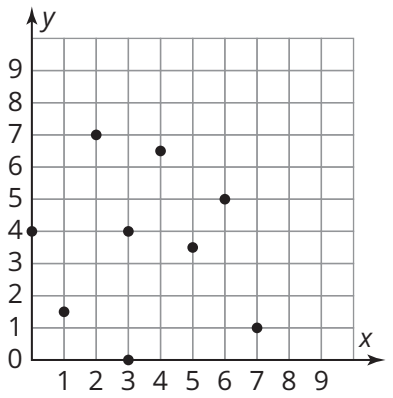
The closer the *r*-value gets to 0, the less of a linear relationship there is in the data.

b.



- $r = 0.7$
- $r = -0.7$
- $r = 0.07$
- $r = -0.07$

c.



- $r = 1$
- $r = 0.5$
- $r = 0.01$

You can calculate the correlation coefficient of a data set using the formula:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Fortunately your graphing calculator can do this arithmetic. Previously you used a graphing calculator to determine the linear regression using the Least Squares Method. Along with calculating the equation for the line, the calculator also calculated the value *r*, the correlation coefficient.

Let's use technology to compute the value of the correlation coefficient.

2. Consider the data set (23, 23), (1, 2), and (3, 4).

a. Use technology to compute the correlation coefficient.

b. Interpret the correlation coefficient of the data set.

ACTIVITY
2.2 Is It Linear?

A group of friends completed a survey about their monthly income and how much they pay for rent each month. The table shows the results.

Monthly Net Income (dollars)	Monthly Rent (dollars)
1400	450
1550	505
2000	545
2600	715
3000	930
3400	1000

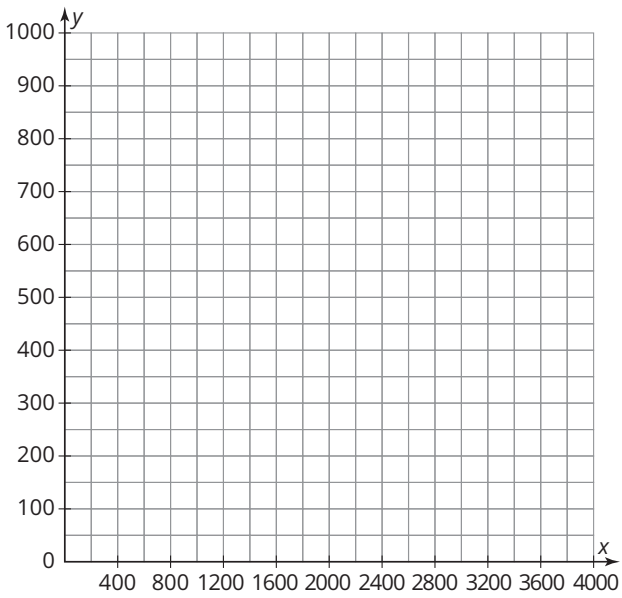
Ask yourself:

What do you notice as you read through the data?

1. Identify the independent and dependent quantities in this problem situation.

2. Construct a scatter plot of the data using technology.

a. Sketch and label the scatter plot.



b. Do you think a linear regression equation would best describe this situation? Explain your reasoning.

Ask yourself:

What is the appropriate level of accuracy needed for this linear regression equation?

3. Use technology to determine whether a line of best fit is appropriate for these data.

a. Determine and interpret the linear regression equation.

b. Compute the correlation coefficient.

4. Would a line of best fit be appropriate for this data set? Explain your reasoning.

The correlation coefficient, *r*, indicates the type (positive or negative) and strength of the relationship that may exist for a given set of data points. The **coefficient of determination**, r^2, measures how well the graph of the regression fits the data. It represents the percentage of variation of the observed values of the data points from their predicted values.

ACTIVITY
2.3

Using the Correlation Coefficient to Assess a Line of Best Fit

The amount of antibiotic that remains in your body over a period of time varies from one drug to the next. The table given shows the amount of Antibiotic X that remains in your body over a period of two days.

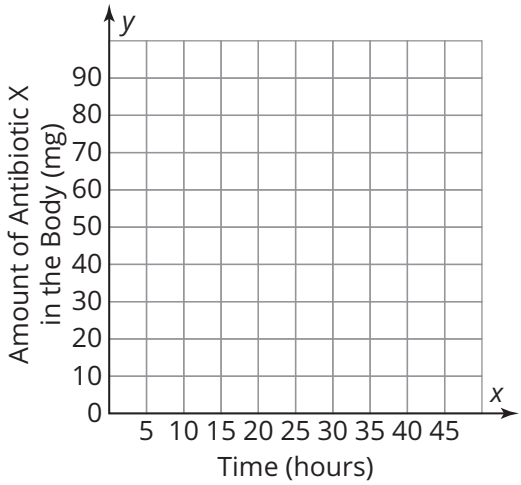
Time (hours)	Amount of Antibiotic X in Body (mg)
0	60
6	36
12	22
18	13
24	7.8
30	4.7
36	2.8
42	1.7
48	1

1. Determine and interpret a linear regression equation for this data set.

2. Compute and interpret both the correlation coefficient and coefficient of determination of this data set.

3. Does it seem appropriate to use a line of best fit? If no, explain your reasoning.

4. Sketch a scatter plot of the data.



5. Look at the graph of the data. Do you still agree with your answer to Question 3? Explain your reasoning.

ACTIVITY
2.4 Correlation Vs. Causation

Does correlation mean causation? What do you think causation means? That is a question that statisticians are always trying to determine.

Read the three true statements that Alonzo and Richard are given by their Algebra I teacher. She asks them to decide what conclusions they can draw from the data. Do you agree with them? If so, why? If not, why not?

1. **The number of smartphones sold in the United States has increased every year since 2005. The number of flat screen televisions sold in the United States has also increased during the same period of time.**

   Alonzo and Richard reached the conclusion that owning a cell phone causes a person to buy a flat screen television.

2. **Since 2004, the average salary of an NFL football player has increased every year. The average weight of an NFL player has also increased yearly since 2004.**

   After much discussion, Alonzo and Richard reached the conclusion that higher salaries cause the players to gain weight.

3. **Worldwide, the number of automobiles sold annually has steadily increased since 1920. Gasoline production has also steadily increased since 1920.**

   Alonzo and Richard concluded that the increase in the number of automobiles sold caused an increase in the amount of gasoline produced.

Proving causation is challenging. The scenarios Alonzo and Richard analyzed demonstrate that even though two quantities are correlated, this does not mean that one quantity caused the other. This is one of the most misunderstood and misapplied uses of statistics.

Causation is when one event effects the outcome of a second event. A correlation is a **necessary condition** for causation, but a correlation is not a **sufficient condition** for causation. While determining a correlation is straightforward, using statistics to establish causation is very difficult.

4. Many medical studies have tried to prove that smoking causes lung cancer.

a. Is smoking a necessary condition for lung cancer? Why or why not?

b. Is smoking a sufficient condition for lung cancer? Why or why not?

c. Is there a correlation between people who smoke and people who get lung cancer? Explain your reasoning.

d. Is it true that smoking causes lung cancer? If so, how was it proven?

5. **It is often said that teenage drivers cause automobile accidents.**

   a. **Is being a teenage driver a necessary condition to have an automobile accident? Why or why not?**

   b. **Is being a teenage driver a sufficient condition to have an automobile accident? Why or why not?**

   c. **Is there a correlation between teenage drivers and automobile accidents? Explain your reasoning.**

   d. **Is it true that teenage drivers cause automobile accidents? Explain your reasoning.**

Does school absenteeism cause poor performance in school? A correlation between the independent variable of days absent to the dependent variable of grades makes sense. However, this alone does not prove causation.

6. **In order to prove that the number of days that a student is absent causes the student to get poor grades, we would need to conduct more controlled experiments.**

   a. **List several ways that you could design experiments to attempt to prove this assertion.**

   b. **Will any of these experiments prove the assertion? Explain your reasoning.**

There are two relationships that are often mistaken for causation. A **common response** is when some other reason may cause the same result. A **confounding variable** is when there are other variables that are unknown or unobserved.

7. Consider each relationship. List two or more common responses that could also cause this result.

a. In North Carolina, the number of shark attacks increases when the temperature increases. Therefore, a temperature increase appears to cause sharks to attack.

b. A company claims that their weight loss pill caused people to lose 20 pounds when following the accompanying exercise program.

NOTES

TALK the TALK

Correlations R Us

Consider the given data sets.

Set A

x	*y*
0	24
2	19
5	12
10	6
20	0

Set B

x	*y*
8	13
10	4
14	15
15	14
19	73

1. Determine the linear regression for each set.

2. Compare the correlation coefficient and the coefficient of determination of each data set. Describe which regression equation is the better fit and why.

Assignment

Write

Complete each sentence.

1. A correlation is a _____ for causation, but a correlation is not a _____ for causation.
2. A _____ is when some other reason may cause the same result.
3. _____ is when one event causes a second event.
4. A _____ is when there are other variables that are unknown or unobserved.
5. The _____ is a value between -1 and 1 that indicates how close the data are to form a straight line.
6. The percentage of variation of the observed values of the data points from their predicted values is represented by the _____.

Remember

Sets of data can frequently be modeled by using a linear function called a regression equation. A value called the correlation coefficient can also be calculated to assist in determining how well the regression equation fits the data.

Practice

1. The table shows the percent of the United States population who did not receive needed dental care services due to cost.

Year	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Percent	7.9	8.1	8.7	8.6	9.2	10.7	10.7	10.8	10.5	12.6	13.3

   a. Do you think a linear regression equation would best describe this situation? Why or why not?
   b. Determine the linear regression equation for these data. Interpret the equation in terms of this problem situation.
   c. Compute and interpret the correlation coefficient of this data set. Does it seem appropriate to use a line of best fit? Explain your reasoning.
   d. Sketch a scatter plot of the data. Then, plot the equation of the regression line on the same grid. Do you still think a linear regression is appropriate? Explain your answer.

2. A teacher claims that students who study will receive good grades.
   a. Do you think that studying is a necessary condition for a student to receive good grades?
   b. Do you think that studying is a sufficient condition for a student to receive good grades?
   c. Do you think that there is a correlation between students who study and students who receive good grades?
   d. Do you think that it is true that studying will cause a student to receive good grades?
   e. List two or more confounding variables that could have an effect on this claim.
3. For each situation, decide whether the correlation implies causation. List reasons why or why not.
   a. The number of violent video games sold in the U.S. is highly correlated to crime rates in real life.
   b. The number of newspapers sold in a city is highly correlated to the number of runs scored by the city's professional baseball team.
   c. The number of mouse traps found in a person's house is highly correlated to the number of mice found in their house.

Stretch

Consider the points: (1, 2), (2, 3), (3, 2), (4, 5), (5, 2.5), (6, 6), (7, 3), (8, 7). The line of best fit for the graph of the points is $y = 0.5x + 1.4$.

1. Complete the table to determine the predicted values of y for each value of x using the line of best fit, and the values of the differences between the observed y-values from the points and the predicted values of y from the line of best fit.

x	Observed Value of y	Predicted Value of y	Observed Value of y − Predicted Value of y
1	2	1.9	0.1
2	3		
3	2		
4	5		
5	2.5		
6	6		
7	3		
8	7		

2. Determine whether there is a pattern in the differences between the y-values from the completed table. Explain what this might indicate about using the line of best fit to make predictions.

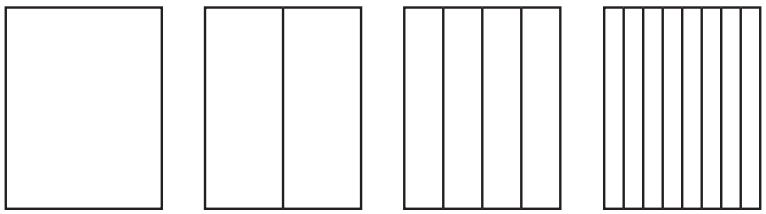
Review

1. The table shows the highest maximum temperature for the month of October in Philadelphia, Pennsylvania, over ten years.

Year	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017
Highest Maximum Temperature (°F)	64.9	53.1	61	54	63	68	61	57.9	64.9	66.9

   a. Identify the independent and dependent quantities and their units of measure.
   b. Use the data table and graphing technology to generate a line of best fit. What is the slope and *y*-intercept of the line and what do they represent?

2. Harrison draws a rectangle, and then in each successive figure he splits the rectangles into two rectangles as shown.



   a. Analyze the number of rectangles in each figure. Describe the pattern.
   b. Write the number of rectangles in each of the first six figures as a numeric sequence.

3. Determine the slope, *x*-intercept, and *y*-intercept of the graph.