Identifying correlation and causation

Sample Question

Answer the questions below.

(a) **A financial magazine conducted a survey. It showed that a person with more years of education tends to have a higher salary. What is likely true?**

- There is *no* correlation between years of education and salary.
- There is a correlation between years of education and salary. There may or may not be causation. Further studies would have to be done to determine this.
- There is a correlation between years of education and salary. There is probably also causation. This is because there is an *increase* in salary with an *increase* in the years of education.

(b) **A grocery store collected sales data. It found that a change in price does not imply that people will buy more or less milk. What can we determine from this information?**

- There is *no* correlation between price and amount of milk bought.
- There is a correlation between price and amount of milk bought. However, there is no causation. This is because there is likely a *decrease* in the amount of milk bought with an *increase* in the price.
- There is a correlation between price and amount of milk bought. There may or may not be causation. Further studies would have to be done to determine this.

(c) **Ms. Bryant studied her students' Spanish test scores and TV habits. She found that students who watched less TV tended to earn higher scores on the test. What conclusion should she make?**

- There is *no* correlation between test score and amount of TV watched.
- There is a correlation between test score and amount of TV watched. There is probably also causation. This is because there is an *increase* in a student's test score with a *decrease* in the amount of TV watched.
- There is a correlation between test score and amount of TV watched. There may or may not be causation. Further studies would have to be done to determine this.

Explanation

Two quantities have a *correlation* if they tend to vary together.

- Suppose there are more bees in a yard when there are more flowers. Then there is a correlation between the number of bees and the number of flowers.
- Suppose less coffee is consumed when the temperature is higher. Then there is a correlation between the amount of coffee consumed and the temperature.

Causation means that a change in one quantity directly causes a change in the other.

Causation can only be shown by a controlled experiment.

Correlation **does not imply** causation.
That is, two quantities can vary together without one of them causing a change in the other.

(a) We see from the following that years of education and salary **tend to vary together.**

"a person with more years of education tends to have a higher salary"

Thus, **there is a correlation** between years of education and salary.

However, we cannot be sure there is causation without further studies.

(b) We see from the following that price and amount of milk bought **tend not to vary together.**

"a change in price does not imply that people will buy more or less milk"

Thus, **there is no correlation** between price and amount of milk bought.

(c) We see from the following that test score and amount of TV watched **tend to vary together.**

"students who watched less TV tended to earn higher scores on the test"

Thus, **there is a correlation** between test score and amount of TV watched.

However, we cannot be sure there is causation without further studies.

Sample Question

Answer the questions below.

(a) **A flute instructor examined student progress. He found that playing a longer song does not necessarily mean that a student will have more or fewer wrong notes. What can he conclude?**

- There is *no* correlation between song length and number of wrong notes.
- There is a correlation between song length and number of wrong notes. There may or may not be causation. Further studies would have to be done to determine this.
- There is a correlation between song length and number of wrong notes. There is probably also causation. This is because there is likely an *increase* in the number of wrong notes with an *increase* in the song length.

(b) **An environmental study was conducted. It reported that, for a park, having more visitors indicates having a larger number of trees. What can be determined?**

- There is *no* correlation between number of visitors and number of trees.
- There is a correlation between number of visitors and number of trees. However, there is no causation. This is because there is an *increase* in the number of trees with an *increase* in the number of visitors.
- There is a correlation between number of visitors and number of trees. There may or may not be causation. Further studies would have to be done to determine this.

(c) **Data was collected at a tennis court. It showed that being older implies that a player will stay for a shorter length of time. What is likely true?**

- There is *no* correlation between age and length of stay.
- There is a correlation between age and length of stay. There may or may not be causation. Further studies would have to be done to determine this.
- There is a correlation between age and length of stay. However, there is no causation. This is because there is a *decrease* in the length of stay with an *increase* in age.

Explanation

Two quantities have a *correlation* if they tend to vary together.

- Suppose there are more bees in a yard when there are more flowers. Then there is a correlation between the number of bees and the number of flowers.
- Suppose less coffee is consumed when the temperature is higher. Then there is a correlation between the amount of coffee consumed and the temperature.

Causation means that a change in one quantity directly causes a change in the other.

Causation can only be shown by a controlled experiment.

Correlation **does not imply** causation.
That is, two quantities can vary together without one of them causing a change in the other.

(a) We see from the following that song length and number of wrong notes **tend not to vary together.**

"playing a longer song does not necessarily mean that a student will have more or fewer wrong notes"

Thus, **there is no correlation** between song length and number of wrong notes.

(b) We see from the following that number of visitors and number of trees **tend to vary together.**

"for a park, having more visitors indicates having a larger number of trees"

Thus, **there is a correlation** between number of visitors and number of trees.

However, we cannot be sure there is causation without further studies.

(c) We see from the following that age and length of stay **tend to vary together.**

"being older implies that a player will stay for a shorter length of time"

Thus, **there is a correlation** between age and length of stay.

However, we cannot be sure there is causation without further studies.

PRACTICE

1.

(a) **A researcher measured the shoe size and reading ability of a large group of children. She found that, as shoe size increases, so does reading ability. What does her analysis show?**

- There is *no* correlation between shoe size and reading ability.
- There is a correlation between shoe size and reading ability. There may or may not be causation. Further studies would have to be done to determine this.
- There is a correlation between shoe size and reading ability. There is probably also causation. This is because there is an *increase* in reading ability with an *increase* in shoe size.

(b) **A grocery store conducted a survey. It showed that when customers buy less cereal, it does not indicate a change in the amount of milk they purchase. What can we conclude?**

- There is *no* correlation between amount of cereal bought and amount of milk purchased.
- There is a correlation between amount of cereal bought and amount of milk purchased. However, there is no causation. This is because there is likely a *decrease* in the amount of milk purchased with a *decrease* in the amount of cereal bought.
- There is a correlation between amount of cereal bought and amount of milk purchased. There may or may not be causation. Further studies would have to be done to determine this.

(c) **Raina compared the player statistics from team's basketball season. She determined that being taller does not imply that a player scores more or fewer points. What is likely true?**

- There is *no* correlation between height and number of points scored.
- There is a correlation between height and number of points scored. There may or may not be causation. Further studies would have to be done to determine this.
- There is a correlation between height and number of points scored. However, there is no causation. This is because there is probably an *increase* in the number points scored with an *increase* in height.

2.

(a) **A health group conducted a survey. It showed that a heavier newborn did not necessarily need a larger or smaller amount of diapers. What is likely true?**

- There is *no* correlation between birth weight and amount of diapers needed.
- There is a correlation between birth weight and amount of diapers needed. There may or may not be causation. Further studies would have to be done to determine this.
- There is a correlation between birth weight and amount of diapers needed. There is probably also causation. This is because there would be a *decrease* in the amount of diapers needed with an *increase* in the birth weight.

(b) **A government study found that people buy more gasoline when the price per gallon decreases. What can we determine?**

- There is *no* correlation between amount of gasoline bought and price.
- There is a correlation between amount of gasoline bought and price. However, there is no causation. This is because there is an *increase* in the amount of gasoline bought with a *decrease* in the price.
- There is a correlation between amount of gasoline bought and price. There may or may not be causation. Further studies would have to be done to determine this.

(c) **Mr. Thompson studied his math students' homework and test scores. He found that students who completed more homework did not tend to earn higher or lower scores on the test. What should he conclude?**

- There is *no* correlation between test score and amount of homework completed.
- There is a correlation between test score and amount of homework completed. There is probably also causation. This is because there might be an *increase* in a student's test score with an *increase* in the amount of homework completed.
- There is a correlation between test score and amount of homework completed. There may or may not be causation. Further studies would have to be done to determine this.

3.

(a) **A golf instructor examined student data. She found that when a student takes longer to hit, it implies that the ball will land closer to the hole. What can she conclude?**

- There is *no* correlation between amount of time before hitting and distance from the hole.
- There is a correlation between amount of time before hitting and distance from the hole. There may or may not be causation. Further studies would have to be done to determine this.
- There is a correlation between amount of time before hitting and distance from the hole. There is probably also causation. This is because there is a *decrease* in the distance from the hole with an *increase* in the amount of time before hitting.

(b) **A city conducted a traffic study. It reported that having more signals does not indicate a shorter or longer commute time for drivers. What can be determined?**

- There is *no* correlation between number of signals and commute time.
- There is a correlation between number of signals and commute time. However, there is no causation. This is because there is probably an *increase* in the commute time with an *increase* in the number of signals.
- There is a correlation between number of signals and commute time. There may or may not be causation. Further studies would have to be done to determine this.

(c) **Employees at a large office building took a survey. The results show that those who began work earlier tended to drink less tea. Which statement is most likely true?**

- There is *no* correlation between start time and amount of tea consumed.
- There is a correlation between start time and amount of tea consumed. There may or may not be causation. Further studies would have to be done to determine this.
- There is a correlation between start time and amount of tea consumed. However, there is no causation. This is because there is a *decrease* in the amount of tea consumed with an earlier start time.

ANSWERS

1. (a) We see from the following that shoe size and reading ability **tend to vary together.**

"as shoe size increases, so does reading ability"

Thus, **there is a correlation** between shoe size and reading ability.

However, we cannot be sure there is causation without further studies.

(b) We see from the following that amount of cereal bought and amount of milk purchased **tend not to vary together.**

"when customers buy less cereal, it does not indicate a change in the amount of milk they purchase"

Thus, **there is no correlation** between amount of cereal bought and amount of milk purchased.

(c) We see from the following that height and number of points scored **tend not to vary together.**

"being taller does not imply that a player scores more or fewer points"

Thus, **there is no correlation** between height and number of points scored.

2. (a) We see from the following that birth weight and amount of diapers needed **tend not to vary together.**

"a heavier newborn did not necessarily need a larger or smaller amount of diapers"

Thus, **there is no correlation** between birth weight and amount of diapers needed.

(b) We see from the following that amount of gasoline bought and price **tend to vary together.**

"people buy more gasoline when the price per gallon decreases"

Thus, **there is a correlation** between amount of gasoline bought and price.

However, we cannot be sure there is causation without further studies.

(c) We see from the following that test score and amount of homework completed **tend not to vary together.**

"students who completed more homework did not tend to earn higher or lower scores on the test"

Thus, **there is no correlation** between test score and amount of homework completed.

3. (a) We see from the following that amount of time before hitting and distance from the hole **tend to vary together.**

"when a student takes longer to hit, it implies that the ball will land closer to the hole"

Thus, **there is a correlation** between amount of time before hitting and distance from the hole.

However, we cannot be sure there is causation without further studies.

(b) We see from the following that number of signals and commute time **tend not to vary together.**

"having more signals does not indicate a shorter or longer commute time for drivers"

Thus, **there is no correlation** between number of signals and commute time.

(c) We see from the following that start time and amount of tea consumed **tend to vary together.**

"those who began work earlier tended to drink less tea"

Thus, **there is a correlation** between start time and amount of tea consumed.

However, we cannot be sure there is causation without further studies.

Module 10: Interpreting Categorical and Quantitative Data

Linear relationship and the correlation coefficient

Sample Question

Shown below are the scatter plots for four different data sets.

Answer the questions that follow. The same response may be the correct answer for more than one question.

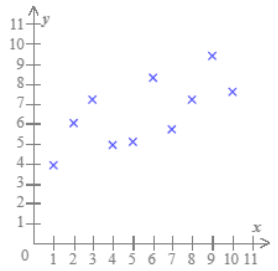


Figure 1

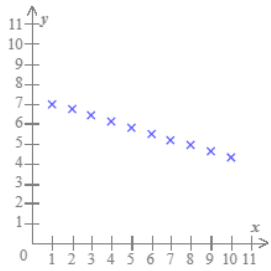


Figure 2

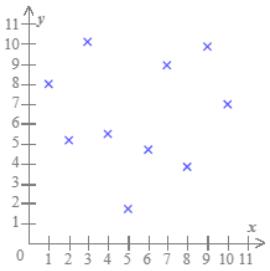


Figure 3

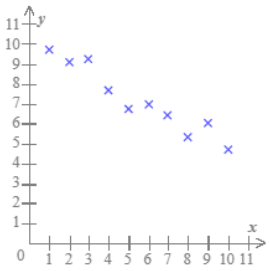


Figure 4

1. Which data set indicates the strongest linear relationship between its two variables?
2. Which data set has an apparent positive, but not perfect, linear relationship between its two variables?
3. For which data set does the correlation coefficient r appear to be equal to -1?

Explanation

Background:

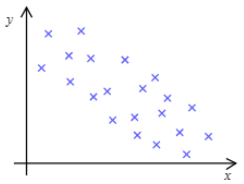
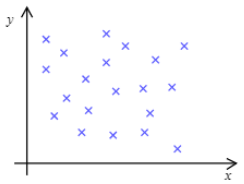
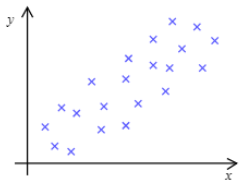
The correlation coefficient, r, measures the linear relationship between two variables.

The value of r is a number from -1 to 1.

A negative value of r indicates a negative linear relationship between the two variables.

A value of r close to 0 indicates there is little or no linear relationship.

A positive value of r indicates a positive linear relationship.

Negative linear relationship	No linear relationship	Positive linear relationship
		
As x increases, y tends to decrease	There is no obvious pattern	As x increases, y tends to increase

The closer the points are to falling on a straight line, the stronger the linear relationship.
The stronger the linear relationship, the closer r is to -1 or 1.

A value of $r = 1$ indicates a perfect positive linear relationship.
The points lie exactly on a straight line that rises from left to right.

A value of $r = -1$ indicates a perfect negative linear relationship.
The points lie exactly on a straight line that falls from left to right.

The current problem:

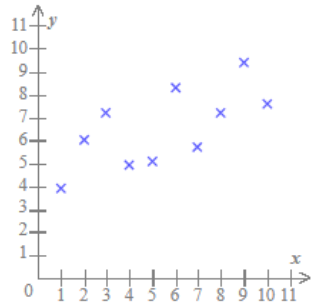


Figure 1

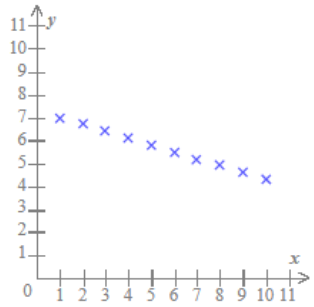


Figure 2

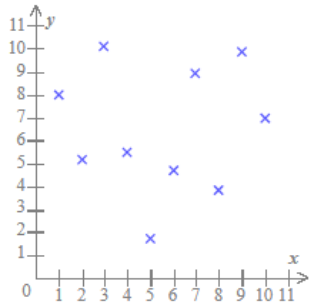


Figure 3

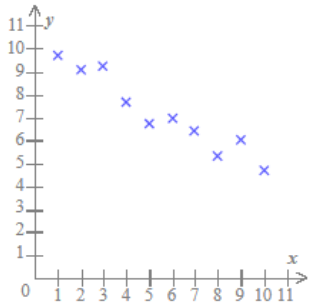


Figure 4

1. **Which data set indicates the strongest linear relationship between its two variables?**

   The data points in Figure 2 lie on a straight line.
   So, this data set has the strongest linear relationship.
   (The value of r is -1 for this data set.)

2. **Which data set has an apparent positive, but not perfect, linear relationship between its two variables?**

   The data points in Figure 1 show a positive, but not perfect, linear relationship.
   (The value of r is approximately 0.68 for this data set.)

3. **For which data set does the correlation coefficient r appear to be equal to -1?**

   The data points in Figure 2 lie exactly on a straight line that falls from left to right.
   So, $r = -1$ for this data set.

Sample Question

Shown below are the scatter plots for four different data sets.

Answer the questions that follow. The same response may be the correct answer for more than one question.

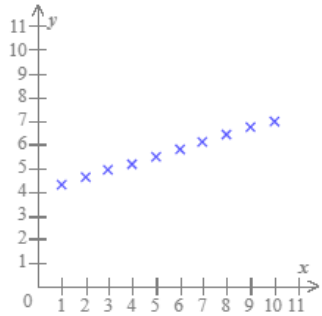


Figure 1

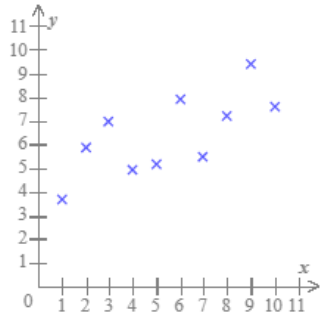


Figure 2

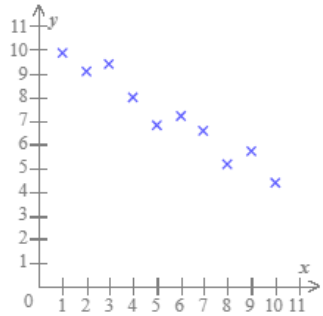


Figure 3

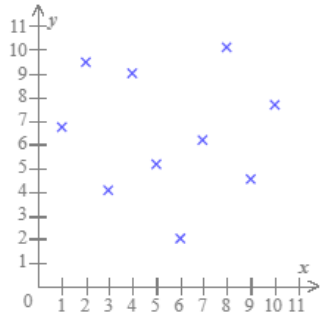


Figure 4

1. For which data set does the correlation coefficient r appear to be equal to 1?
2. Which data set shows the least evidence of a linear relationship?
3. For which data set is the correlation coefficient r closest to -1?

Explanation

Background:

The correlation coefficient, r, measures the linear relationship between two variables.

The value of r is a number from -1 to 1.

The current problem:

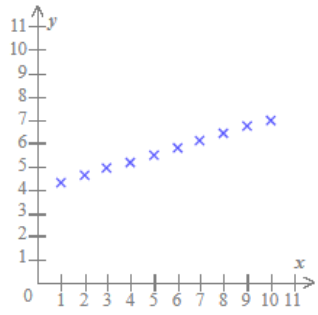


Figure 1

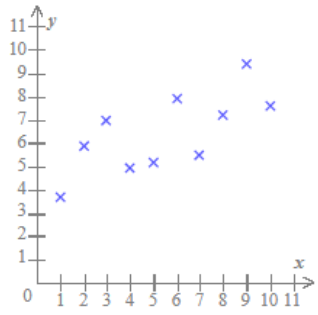


Figure 2

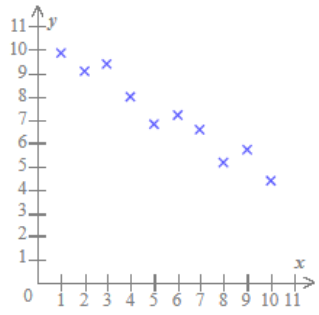


Figure 3

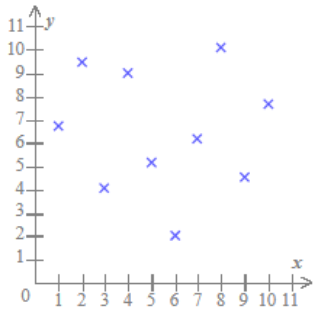


Figure 4

1. **For which data set does the correlation coefficient r appear to be equal to 1?**

   The data points in Figure 1 lie exactly on a straight line rises from left to right.
   So, $r = 1$ for this data set.

2. **Which data set shows the least evidence of a linear relationship?**

   Figure 4 shows the only data set with no obvious pattern.
   So, this data set shows the least evidence of a linear relationship.
   (The value of r is approximately -0.05 for this data set.)

3. **For which data set is the correlation coefficient r closest to -1?**

   The only data set that shows a negative relationship is the one in Figure 3.
   So, this data set has a value of r closest to -1.
   (The value of r is approximately -0.97 for this data set.)

PRACTICE

1. Shown below are the scatter plots for four different data sets.

Answer the questions that follow. The same response may be the correct answer for more than one question.

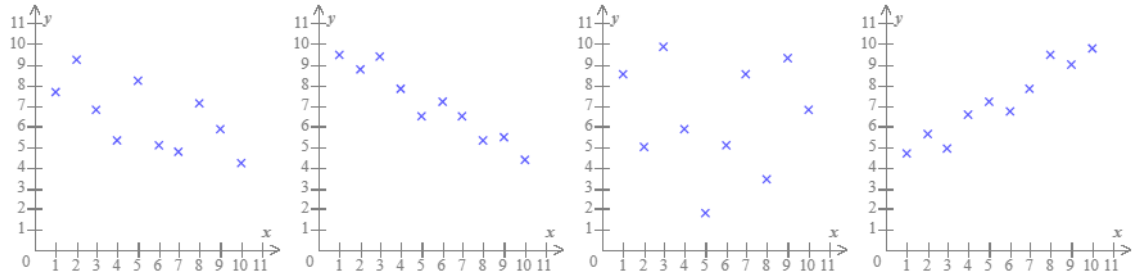


Figure 1 | Figure 2 | Figure 3 | Figure 4

1. For which data set is the correlation coefficient ***r*** closest to 1?

_____

2. Which data set indicates the strongest negative linear relationship between its two variables?

_____

3. For which data set is the correlation coefficient ***r*** closest to 0?

2. Shown below are the scatter plots for four different data sets.

Answer the questions that follow. The same response may be the correct answer for more than one question.

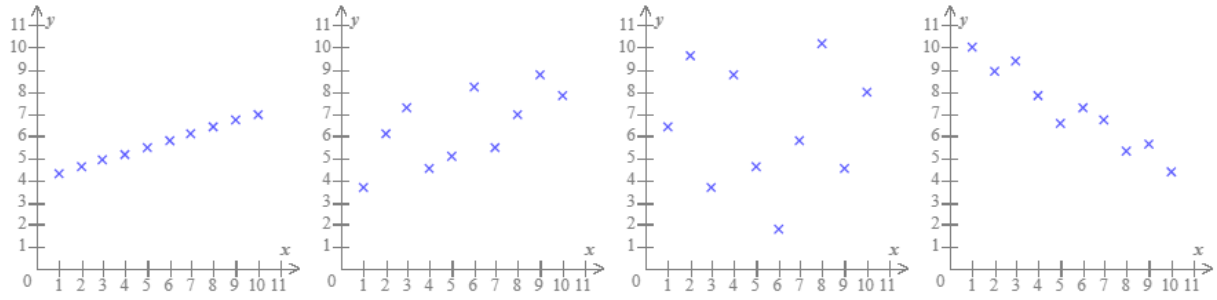


Figure 1 | Figure 2 | Figure 3 | Figure 4

1. Which data set indicates the strongest linear relationship between its two variables?

_____

2. Which data set has an apparent positive, but not perfect, linear relationship between its two variables?

_____

3. For which data set does the correlation coefficient ***r*** appear to be equal to 1?

ANSWERS

1. 1. For which data set is the correlation coefficient r closest to 1?

The only data set that shows a positive linear relationship is the one in Figure 4.
So, this data set has a value of r closest to 1.
(The value of r is approximately 0.96 for this data set.)

2. Which data set indicates the strongest negative linear relationship between its two variables?

The data sets in Figure 1 and Figure 2 both show negative linear relationships.
Of the two data sets, the points in Figure 2 are closer to lying on a straight line.

So, the data set in Figure 2 shows the strongest negative linear relationship.
(The values of r are approximately -0.65 and -0.96 for the data sets in Figures 1 and 2.)

3. For which data set is the correlation coefficient r closest to 0?

Figure 3 shows the only data set with no obvious pattern.
So, this data set has a value of r closest to 0.
(The value of r is approximately -0.05 for this data set.)

2. 1. Which data set indicates the strongest linear relationship between its two variables?

The data points in Figure 1 lie on a straight line.
So, this data set has the strongest linear relationship.
(The value of r is 1 for this data set.)

2. Which data set has an apparent positive, but not perfect, linear relationship between its two variables?

The data points in Figure 2 show a positive, but not perfect, linear relationship.
(The value of r is approximately 0.66 for this data set.)

3. For which data set does the correlation coefficient r appear to be equal to 1?

The data points in Figure 1 lie exactly on a straight line that rises from left to right.
So, $r = 1$ for this data set.