Unit 3

Describing Data

In this unit, you will use dot plots, histograms, box plots, and measures of center and spread to analyze and describe one-variable data sets. You will also use scatter plots, correlation coefficients, and lines of best fit to describe two-variable data sets.

Essential Questions

- What tools can you use to help determine if there is an association in a set of data?
- How can you compare data using measures of center and spread?
- How can you use the correlation coefficient and line of best fit to describe the relationship between two variables?

Summary | Lesson 1

Different types of questions lead to different types of data:

- **Quantitative data** has values that are numbers, measurements, or quantities instead of words. It's sometimes called *numerical data*.

  How many pets do you have? is a question that produces quantitative data.

- **Categorical data** has values that are categories, such as colors, words, or zip codes.

  What's your favorite animal? is a question that produces categorical data.

It's important to be specific when writing survey questions, so you can gather the exact type of data you need.

Try This

Determine whether each survey question will produce categorical or quantitative data.

a. How many languages do you speak?

b. Are you left-handed or right-handed?

c. What is your height?

d. What is your phone number?

Summary | Lesson 2

It can be helpful to represent categorical data in a **two-way table**. Two-way tables show the distribution of a population according to two categorical variables, including totals for each sub-group and for the population as a whole.

A **relative frequency table** shows the same data as percents relative to the totals for one of the variables. You can use this representation to see if the data presents evidence that there is an **association** between the two variables.

In these tables, there is evidence of an association in the data between age and whether a student takes dance class. This association is suggested because the students under 13 are much more likely (≈58.9%) to take dance class than the students 13 or older (≈46.5%).

Arts Studio Students

	Takes Dance Class	Does Not Take Dance Class	Total
13 or Older	105	121	226
Under 13	89	62	151
Total	194	183	377

Arts Studio Students

	Takes Dance Class	Does Not Take Dance Class	Total
13 or Older	≈46.5%	≈53.5%	100.0%
Under 13	≈58.9%	≈41.1%	100.0%

Try This

Here are two tables that show some athletes' states of mind during a track meet and whether or not they meditated beforehand.

a Complete each table.

Two-Way Table

	Meditated	Did Not Meditate	Total
Calm	45		53
Anxious		21	
Total	68	29	97

Relative Frequency Table

	Meditated	Did Not Meditate	Total
Calm	≈85%		100%
Anxious			100%

b Choose *one* number from each table and explain what it means.

Summary | Lesson 3

Two-way tables can be helpful when making a decision. The distribution of data in a two-way table can indicate whether there are associations in the data or other information that might support you in making a fair and informed decision.

For example, compare this relative frequency table and its corresponding two-way table. The tables show the data gathered by Tyani's principal about which field trip the students would prefer to take: going whale watching or going to an art museum.

The relative frequency table shows that Tyani's homeroom has a stronger preference for whale watching than the other homerooms, but the two-way table shows that there are more students overall who would prefer to go to the art museum.

Relative Frequency Table

	Whale Watching	Art Museum	Total
Tyani's Homeroom	≈61.5%	≈38.5%	100.0%
Other Homerooms	≈46.1%	≈53.9%	100.0%

Two-Way Table

	Whale Watching	Art Museum	Total
Tyani's Homeroom	16	10	26
Other Homerooms	88	103	191
Total	104	113	217

Try This

This table shows some athletes' states of mind during a track meet and whether or not they meditated beforehand.

	Meditated	Did Not Meditate
Calm	≈66%	
Anxious		≈72%
Total	100%	100%

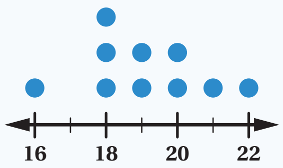
a Complete the relative frequency table.

b Based on this data, is there an association between meditating and state of mind during a track meet? Explain your thinking.

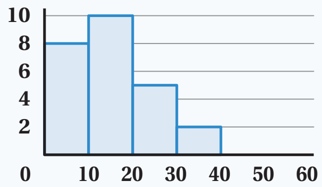
Summary | Lesson 4

It can be helpful to visualize data that's presented in lists or tables using dot plots or histograms.

Dot plots present each data point as a dot at its value on a number line. Dots with the same value are stacked on top of one another.



Histograms group data into rectangular bins. The height of each rectangle shows how many values are in that bin.



A dot plot is useful for observing the frequency of individual data, and the number of points in a data set. Histograms are useful for representing large data sets and observing the shape of a distribution.

Try This

Here is a histogram of how students rated the season fall/autumn on a scale from 0–10.

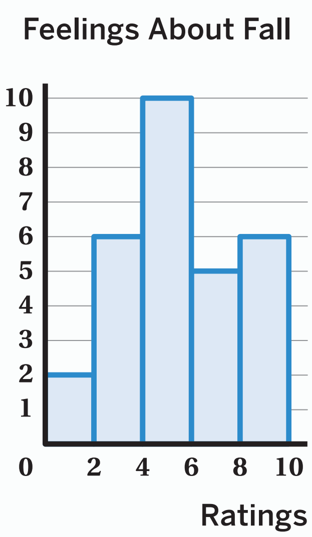
Determine whether each statement is true, false, or cannot be determined using the histogram.

a. There are 29 total ratings.

b. The highest rating was a 9.

c. The lowest rating was less than 2.

d. There are 10 ratings of 6 or higher.

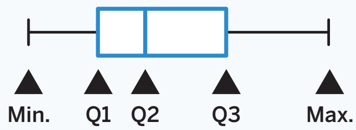


Summary | Lesson 5

A **box plot** is one way to visualize quantitative data. The data is divided into four sections using five values: the *minimum*, the *maximum*, and three quartiles. *Quartiles* divide a data set into four sections, or quarters. Quartile 1 (Q1) is the median of the lower half of the data. Quartile 2 (Q2) is also the *median* of the entire data set, which divides the data into two halves. Quartile 3 (Q3) is the median of the upper half of the data.

The minimum, maximum, median, and quartiles are all examples of statistics. A **statistic** is a single number that measures something about a data set.

To make a box plot:



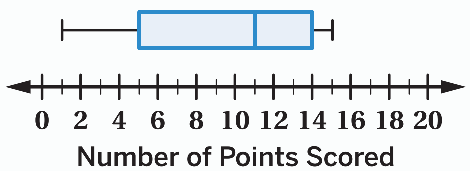
- Calculate the values for Q1, Q2, and Q3.
- Mark the values for the minimum, maximum, Q1, Q2, and Q3.
- Draw a box between Q1 and Q3.
- Draw a vertical line inside the box at Q2 to represent the median.
- Draw horizontal lines from the edges of the box to both the minimum and maximum values.

Box plots are useful for representing large data sets, especially those with extreme values. They are less helpful for seeing the shape of the data.

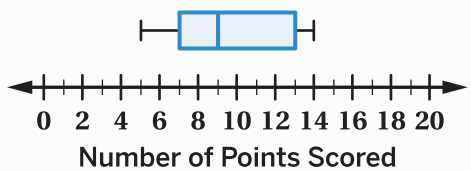
Try This

Two basketball players recorded their points for each game in the season.

Player A



Player B



Determine whether each statement is true, false, or cannot be determined using the box plots.

a The median of Player B's data is 9 points.

b Player A played 15 games this season.

c Player A scored 0 points in at least one game.

d Player B scored between 7 and 9 points in about the same number of games as they scored between 9 and 13 points.

Summary | Lesson 6

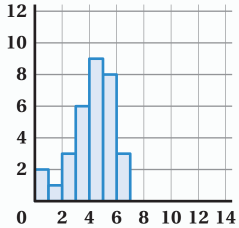
Here are some terms that can help us describe the **shape** of a data set.

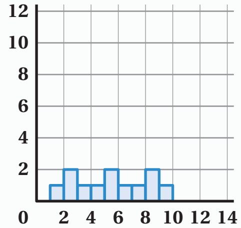
- A data set is **bell-shaped** when most of the data is at the center and there are fewer points farther from the center. When presented in a dot plot or histogram, the data looks like a bell.
- A data set is **bimodal** when it has two very common data values. These appear in a dot plot or histogram as two peaks.
- A data set is **skewed** when more values are concentrated on one end of the data than the other.
- A data set is **symmetric** when you can draw a vertical line of symmetry through it.
- A data set is **uniform** when the data values are evenly distributed.

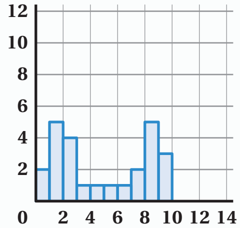
Try This

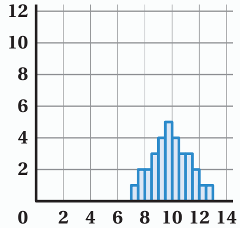
Match each histogram with the best description of its shape.

Bimodal	Bell-shaped	Skewed	Symmetric

a. 

b. 

c. 

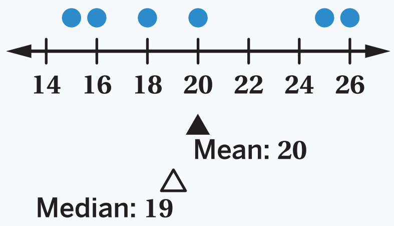
d. 

Summary | Lesson 7

A **measure of center** is a single number that summarizes all of the values in a data set. *Mean* and *median* are measures of center that are used to describe a typical value of a data set.

- The mean is also called the average of a data set. To calculate the mean, you can add up all the data values, and divide by the number of data points.
- The median is the middle value of a data set when the values are in numerical order. If there are two values in the middle of the data set, then the median is the middle of those two values.

Here are the steps you can use to determine the mean of the data in this dot plot.



$15 + 16 + 18 + 20 + 25 + 26 = 120$

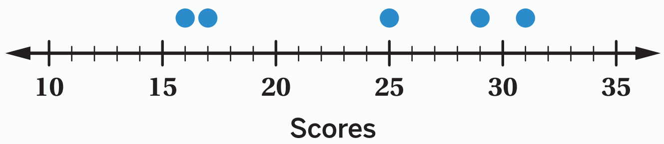
$120 \div 6 = 20$

The median is the value halfway between 18 and 20, so the median is 19.

When a data set includes extreme values that are much larger or smaller than most of the data, the value of the mean and median will be very different. Extreme values impact the mean more than they impact the median.

Try This

Here is a dot plot of Oscar's scores from his newest video game.



a Calculate the mean and median of Oscar's data.

b Oscar plotted his last score incorrectly. He actually scored a 6 instead of a 16.

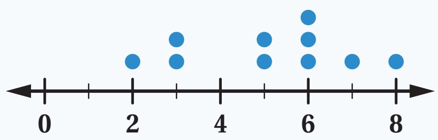
Which measure(s) of center would be affected by this mistake? Explain your thinking.

Summary | Lesson 8

The **standard deviation** of a data set is a **measure of spread**, or a way to measure how spread out the values in a data set are from its mean. Measures of spread can help you describe how consistent a data set is. If a data set is very consistent and much of the data clustered together, the standard deviation will be small. If the data is spread out, the standard deviation will be larger.

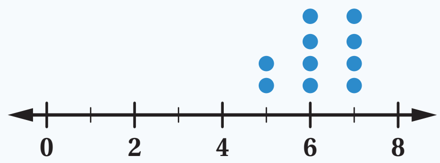
For example, compare these two dot plots.

Dot Plot A



Standard deviation: 1.81

Dot Plot B



Standard deviation: 0.75

The data in Dot Plot A has a greater standard deviation than the data in Dot Plot B because the data in A is more spread out, or variable. Since the data values in Dot Plot B are more consistent, B has a lower standard deviation.

Try This

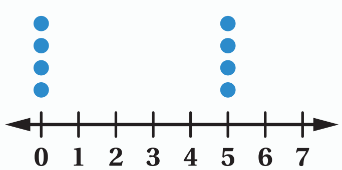
Here are three data sets.

a Which data set do you think has the lowest standard deviation?

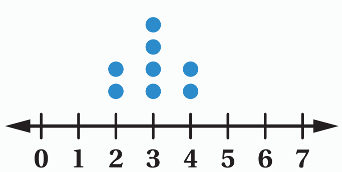
Explain your thinking.

b Use a graphing calculator and the Unit 3 Calculator Guide to calculate the mean and standard deviation for each data set.

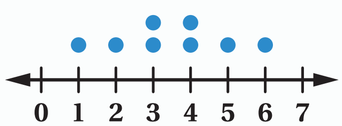
Data Set A



Data Set B



Data Set C



Summary | Lesson 9

You can use statistics, like *mean*, *median*, or standard deviation, to help you compare data sets and make conclusions about the real world. Different statistics can reveal different kinds of information about the data sets you are comparing. Comparing the mean of two data sets can help you determine which one generally has higher or lower values. Comparing the standard deviation of two data sets can help you determine which one is more consistent.

For example, here are statistics about the high temperatures in Metropolis and Springtown over one week.

	Mean (°F)	Standard Deviation (°F)
Metropolis	74	6
Springtown	74.43	2.06

When you compare the means, you can see that both Metropolis and Springtown had similar high temperatures in that week. But when you compare the standard deviations, you can see that the temperature in Springtown was more consistent than the temperature in Metropolis because the data set has a lower standard deviation

Try This

Three people travel to work each day. Here are the means and standard deviations for each of their travel times.

Person A	Person B	Person C
Mean: 29.71 minutes Standard Deviation: 2.60 minutes	Mean: 38.71 minutes Standard Deviation: 3.10 minutes	Mean: 28.14 minutes Standard Deviation: 8.68 minutes

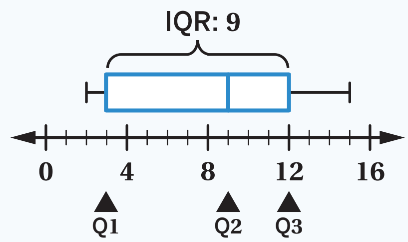
a Which commuter's travel times were most consistent? Use statistics to justify your claim.

b Which commuter had the longest commute time? Use statistics to justify your claim.

Summary | Lesson 10

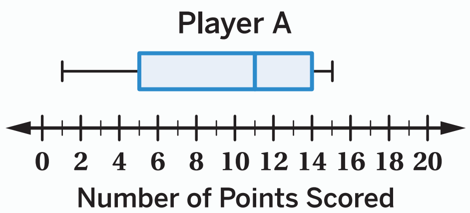
The *interquartile range* (IQR) is a statistic that measures the spread of a data set. Measures of spread, like IQR, help us determine how consistent the data within a set is. The IQR represents the middle half of the data set, and we calculate it by determining the distance from Q1 to Q3. In a box plot, the IQR is the width of the box.

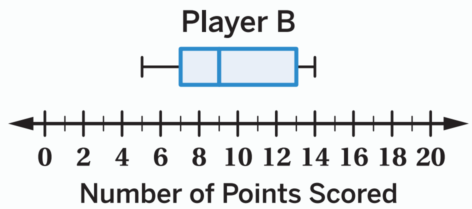
Here is an example of a data set where Q1 $= 3$ and Q3 $= 12$. The IQR of this data set is 9, because $12 - 3 = 9$.



Try This

Two basketball players recorded their points for each game in the season.





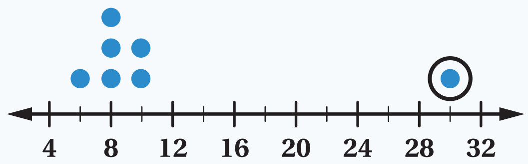
a Determine the median and interquartile range (IQR) for each player.

b Which player was more consistent in their points scored? Use statistics to justify your claim.

c Which player generally scored more points? Use statistics to justify your claim.

Summary | Lesson 11

Outliers are data values that are far from the other values in the data set. In this data set, the circled data point is an outlier.



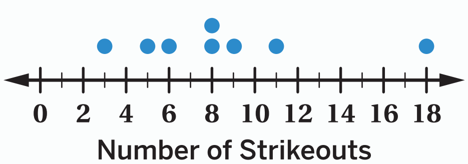
You can identify outliers using dot plots, box plots, and technology tools such as graphing calculators. You can also identify outliers using the *IQR*. Outliers are values further than 1.5 times the IQR below Q1 or above Q3.

When deciding which measure of center is appropriate to represent a data set, it's important to identify any outliers. Outliers have a big impact on the *mean*, but they don't impact the *median* very much.

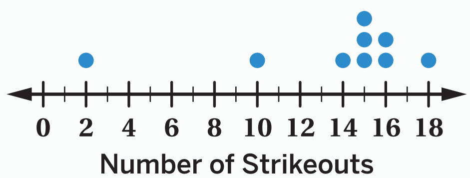
Try This

Let's compare the pitchers on two different baseball teams. These dot plots show the number of strikeouts each pitcher threw in the recent games.

Pitcher A



Pitcher B



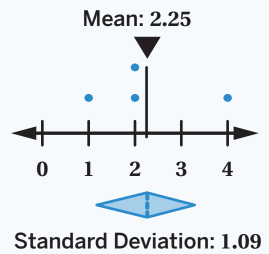
a Use a graphing calculator to make a box plot and identify any outliers in each data set.

b If you removed any outliers from each data set, what would be more affected, the mean or the median? Explain your thinking.

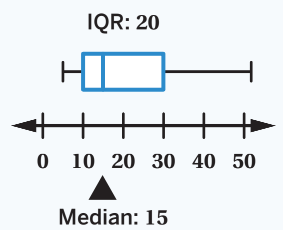
Summary | Lesson 12

It can be helpful to use measures of center and measures of spread to compare data sets. You can choose which measure of center or spread to use based on the shape of the data:

- When data distributions are symmetric or bell-shaped, you can use the mean and standard deviation to compare.
- When data distributions are skewed or contain outliers, you can use the median and IQR because the mean and standard deviation are both affected by extreme values.

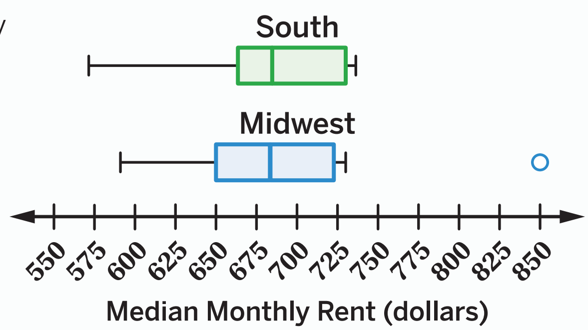


Comparing different data sets can reveal important information about change over time in different situations, like minimum wage or median rent. Comparing measures of center, like mean or median, can help you determine if the data has increased, decreased, or stayed the same. Comparing measures of spread, like standard deviation or IQR, can help you determine if the data has become more or less consistent over time.



Try This

Here is a box plot showing the median monthly rent for eight states in the South and Midwest in 2019.



a What measures of center and spread would you use to compare the rents in each group of states?

Explain your thinking.

b How did rents in the Midwest compare to rents in the South in 2019?

Use statistics to justify your claims.

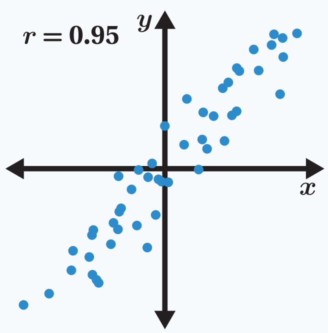
Summary | Lesson 13

A **scatter plot** is a graph of plotted points that shows the relationship, or association, between two variables. When there is a *linear association* between the variables, you can describe the *strength* and *direction* of the association using the **correlation coefficient** (also called the **r-value**). The correlation coefficient is a number between -1 and 1 that describes the strength and direction of a linear association between two variables in a scatter plot.

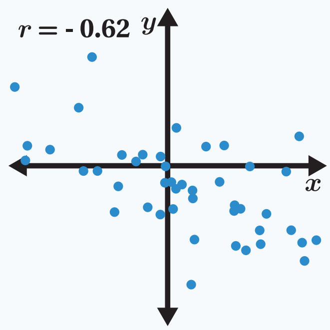
The closer the r-value is to 0, the weaker the linear association. The closer the r value is to -1 or 1, the stronger the linear association.

These graphs show examples of different correlation coefficients.

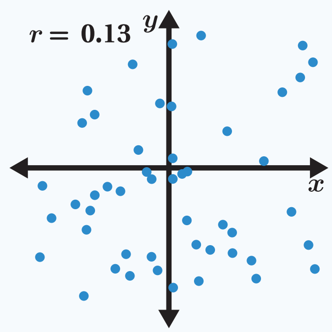
Positive Association
Strong Association



Negative Association



Positive Association
Weak Association



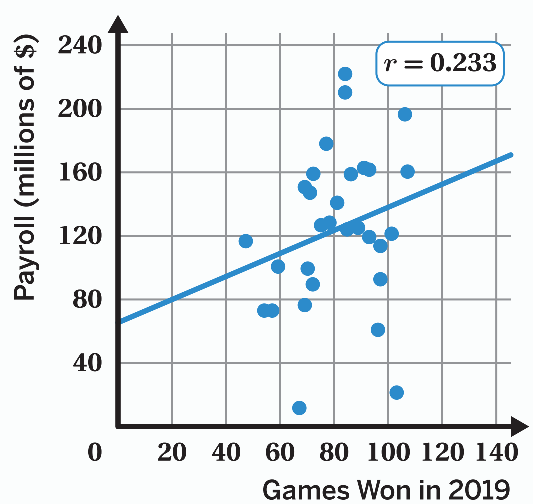
Try This

Lucy was curious about the relationship between how often baseball teams win games and how much money they spend on their players (called payroll).

She found data about games won in 2019 and the team's payroll (in millions of dollars).

a What is the correlation coefficient for Lucy's data?

b Use the correlation coefficient to describe the strength and direction of the association between wins in 2019 and payroll.



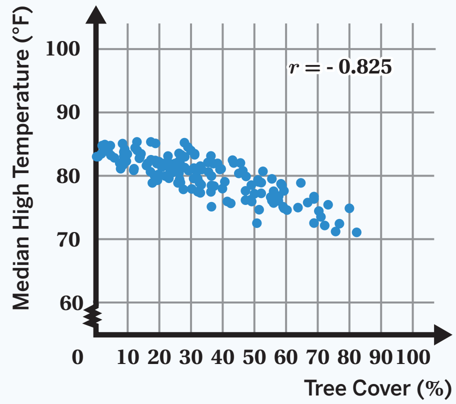
Summary | Lesson 14

You can use a correlation coefficient to analyze the relationship between two variables and determine whether there is an association between them. The correlation coefficient, or r-value, describes the strength and direction of the relationship that may exist between two variables.

- A positive r-value means that as one variable increases, the other variable also increases.
- A negative r-value means that as one variable increases, the other variable decreases.
- The closer the r-value is to 1 or -1, the stronger the correlation.

People may use correlations in data to understand and address issues in their community.

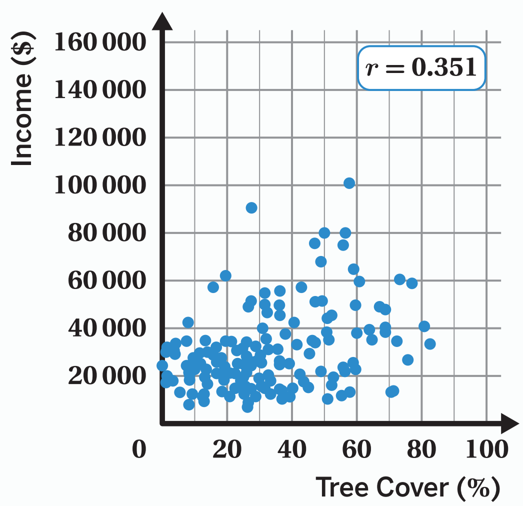
For example, this scatter plot shows data on tree cover and temperature for 150 blocks in Detroit, Michigan.



The r-value is -0.825. This means there is a negative and strong relationship between the amount of tree cover and median high temperature in Detroit neighborhoods. Community members may use this correlation to advocate for more trees to be planted in different neighborhoods across the city.

Try This

Here is a scatter plot with data about the percentage of tree cover recorded in an area of Detroit, Michigan and the income of residents who live there.



a Use the correlation coefficient to describe the association between tree cover and income.

b How might someone use this data and r-value to advocate for their community?

While the correlation coefficient can help you understand the general relationship between two variables, a *line of fit* can help you make predictions about specific values in a data set. Points along the line of fit represent the likely value of unknown data in the data set.

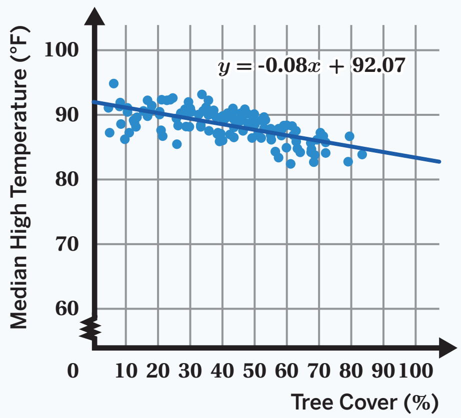
You can use the equation of the line of fit to better understand the data. The slope of the line represents a potential initial value and the y-intercept describes the rate that the variables change in relationship to one another.

For example, this scatter plot shows data on tree cover and temperature for 150 blocks in Austin, Texas. The equation of the line of fit is $y = -0.08x + 92.07$.

The slope is -0.08. This means that when the tree cover increases by 1% in Austin, the predicted temperature decreases by 0.08°F.

The y-intercept is 92.07. This means that if the tree cover in Austin is 0%, the predicted temperature is 92.07°F.

You can use the line to predict that if a block in Austin has 80% tree cover, the temperature will be about 85°F.



Try This

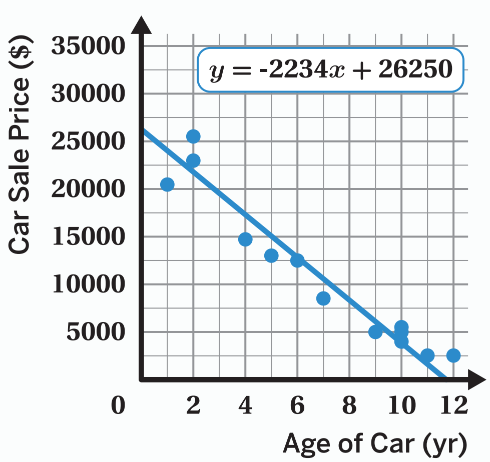
Kwasi was curious about the relationship between the ages of cars and their values. He found data on the ages of several cars (in years) and their sale prices (in dollars) and created a line of fit.

a What does each value represent?

-2,234:

26,250:

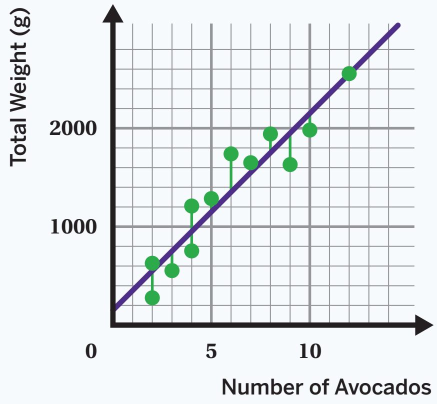
b If a car is 3 years old, what does the line of fit predict its price will be?



Summary | Lesson 16

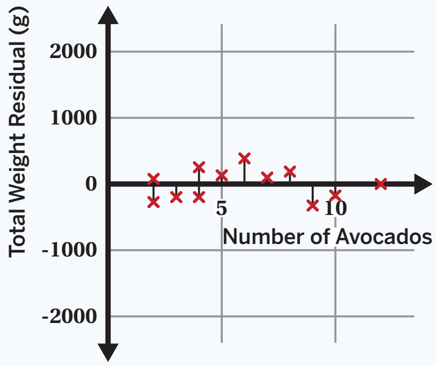
You can use residuals to determine how well a line fits a data set. A **residual** is the difference between the y-value for a point in a scatter plot and the value predicted by the line of best fit.

Here is a scatter plot with data on the number of avocados and their weights. The residuals are represented with lines connecting each point to the line of best fit.



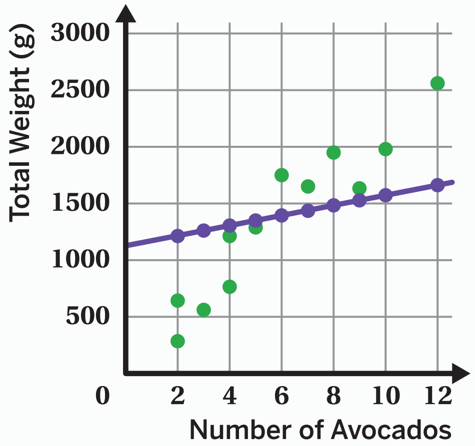
You can also create a residual plot to analyze how well a line fits a data set. A **residual plot** is a scatter plot of residual values for a data set where the x-axis represents the value predicted by the line of best fit, and the y-value of each point represents the value of the residual.

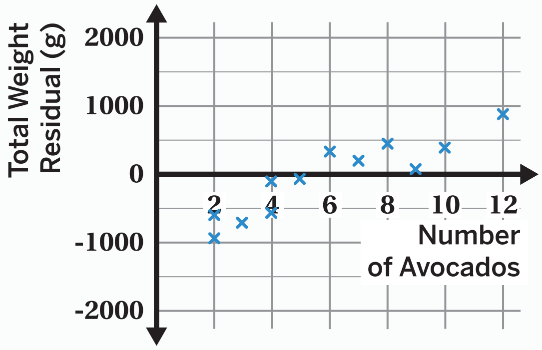
Here is the residual plot of the graph of avocado weights. The closer a point is to the x-axis, the closer that point is to the line of best fit.



Try This

Here is a line of fit for some data about avocados and the residual plot for the line of fit.





a Explain why the point on the residual plot that represents 8 avocados is positive and the point for 3 avocados is negative.

b Use the residual plot to explain why this line is *not* a good fit for the data.

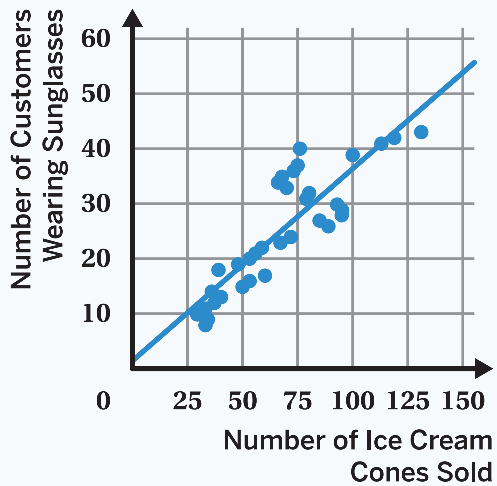
The **line of best** fit is the line on a scatter plot that best represents the trend created by the points in a data set.

Instead of sketching a line of fit, you can use a graphing calculator to precisely generate the equation of the line of best fit from a scatter plot. The equation of the line of best fit can help you interpret information about a situation, or allow you to substitute values into the equation to make predictions. You can also use a graphing calculator to calculate the correlation coefficient for a given data set.

Here is an example of a line of best fit generated by a graphing calculator, and the information about the line that a graphing calculator will show you.

$y_1 \sim mx_1 + b$

STATISTICS	PARAMETERS
$r_2 = 0.7642$	$m = 0.351312$
$r = 0.8742$	$b = 1.31984$

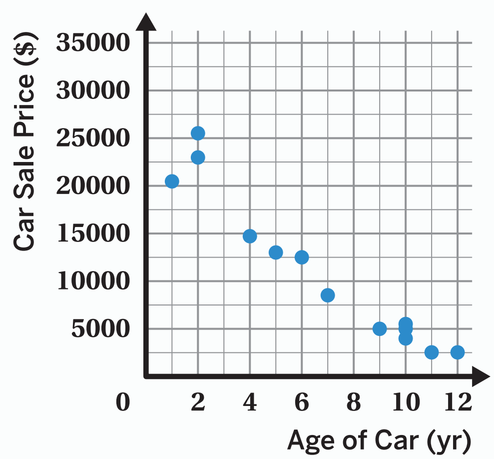


Try This

Kwasi was curious about the relationship between the ages of cars and their values. He found data on the ages of several cars (in years) and their sale prices (in dollars).

a Use the information from the graphing calculator to determine the equation of the line of best fit and the correlation coefficient.

b What does the correlation coefficient say about the association between car age and sale price?



$y_1 \sim mx_1 + b$

STATISTICS	PARAMETERS
$r^2 = 0.9215$	$m = -2270.38$
$r = -0.96$	$b = 26886.7$

Summary | Lesson 18

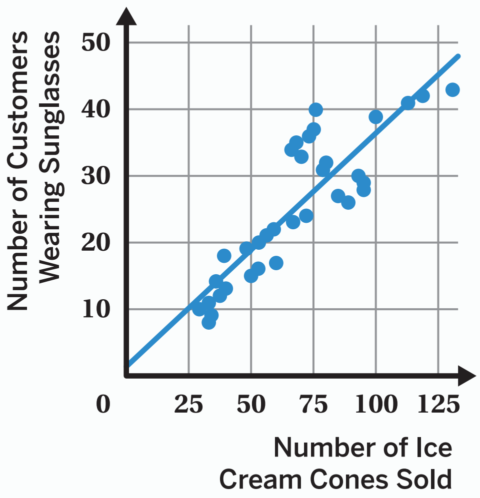
A **correlation**, sometimes called an association, is a relationship between two or more variables. One type of correlation is **causation**, which describes a relationship where a change in one variable causes a change in the other variable.

But not every correlation is a *causal relationship*. For example, a third variable can cause the relationship between two variables to change. Correlation can even be caused by coincidence: if lots of variables are considered, then it's very likely that at least two of them will be somewhat correlated.

Media, such as article headlines, might present relationships that are correlated as causal. To be a critical consumer of information, check the sources behind claims and ask questions about what conclusions can be made from a data set.

Try This

Nyanna noticed a trend at an ice cream shop. She recorded the number of ice cream cones sold and the number of customers wearing sunglasses one day. Then she calculated the r-value, which was 0.87.



a What does the 0.87 say about the association between these two variables?

b Do you think one of the variables causes the other? If not, what else could be affecting the relationship?

Explain your thinking.

Summary | Lesson 19

We can use statistical tools to help us better understand our communities and the world. Scatter plots help us visualize large amounts of data, lines of best fit help us see trends and relationships or associations in that data, and correlation coefficients (r-values) help us describe the strength and direction of any relationships that we find.

While all of these tools help us describe associations and test predictions, we need to explore the variables and relationships further to determine things like the cause and the effect on real people. Statistical tools are powerful, but they are just one step in understanding issues that affect real communities.

Try This

a What are two variables you could explore the association between using the statistics tools in this unit?

b Write a question about your variables that you could answer using the statistics tools from this unit.

c Write a question about your variables that you could *not* answer using the statistics tools from this unit.

Lesson 1

a. Quantitative data

b. Categorical data

c. Quantitative data

d. Categorical data. [Even though these are numbers, phone numbers are not measurements or quantities (e.g., 123-456-7890 is not less than 234-567-8901 in any meaningful way), so this is categorical data.]

Lesson 2

a.

Two-Way Table

	Meditated	Did Not Meditate	Total
Calm	45	**8**	53
Anxious	**23**	21	**44**
Total	68	29	97

Relative Frequency Table

	Meditated	Did Not Meditate	Total
Calm	≈ 85%	**≈ 15%**	100%
Anxious	**≈ 52%**	**≈ 48%**	100%

[One strategy for calculating the missing values in the relative frequency table is to divide each value in the two-way table by the total for the row. For example, 23 people meditated out of the 44 total anxious people, so the percentage of anxious people who meditated is $\frac{23}{44}$ or ≈ 52%.]

b. *Responses vary.*

- 23 is the number of athletes who meditated before the meet and were anxious during the meet.
- ≈ 52% of all the athletes who were anxious meditated before the meet.

Lesson 3

a. Responses shown in the table.

b. Yes. *Explanations vary.* The group that meditated had a much lower percentage of athletes who were anxious. More than half of the athletes who meditated were calm, but only about a quarter of the athletes who did not meditate were calm.

	Meditated	Did Not Meditate
Calm	≈ 66%	**≈ 28%**
Anxious	**≈ 34%**	≈ 72%
Total	100%	100%

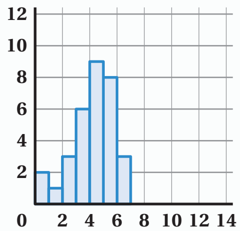
Lesson 4

- **a** True. [One strategy for determining the total number of ratings is to add the ratings in each bin. $2 + 6 + 10 + 5 + 6 = 29$ ratings.]
- **b** Cannot be determined. [The histogram shows that the highest rating was between 8 and 10 but does not say precisely what the rating is.]
- **c** True. [The bin with the lowest ratings has ratings between 0 and 2.]
- **d** False. [There are 11 ratings higher than 6.]

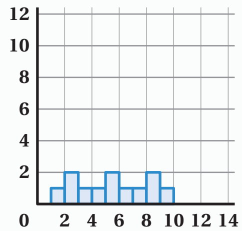
Lesson 5

- **a** True. [The median is the value of quartile 2. The box plot shows that the median for Player B is 9 points.]
- **b** Cannot be determined. [Box plots do not show the number of data points, only groups of 25% of the data.]
- **c** False. [The minimum for Player A is 1 point.]
- **d** True, or cannot be determined. [On a box plot, each section represents about 25% of the data. Even though the section between 9 and 13 points looks larger than the section between 7 and 9 points, they each include about one quarter of the total games. Depending on the exact number of games, they may be slightly different, but that information isn't shown in the box plot.]

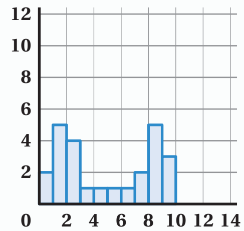
Lesson 6

a. 

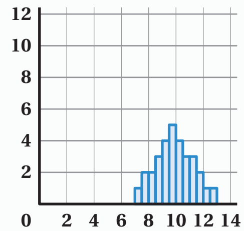
Skewed

b. 

Symmetric

c. 

Bimodal

d. 

Bell-shaped

Lesson 7

a. Mean: 23.6, Median: 25

b. The mean would be affected. *Explanations vary*. Every value is used when calculating the mean, so changing any value affects the mean. The median is not affected by the change because 25 is still the middle value.

Lesson 8

a. Data Set B. *Explanations vary*. The data points in Set B are closest together, so the variation between them is lowest and the standard deviation is lowest.

b. Data Set A: Mean: 2.5, Standard Deviation: 2.5

Data Set B: Mean: 3, Standard Deviation: 0.7

Data Set C: Mean: 3.5, Standard Deviation: 1.5

Note: The Calculator Guide includes instructions on how to use the Desmos Graphing Calculator to calculate the mean and standard deviation of a data set. Use the stdevp( ) function in the calculator because each of these data sets is a population rather than a sample.

Lesson 9

a. Person A. *Explanations vary*. Person A's travel times are the most consistent because they have the smallest standard deviation.

b. Person B. *Explanations vary*. I think Person B has the longest commute time because their mean travel time is the longest.

Lesson 10

a. Player A: Median: 11, IQR: 9

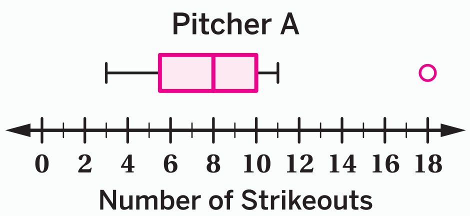
Player B: Median: 9, IQR: 6

b. Player B. *Explanations vary*. Player B is more consistent because their data has a smaller IQR, which is a measure of spread.

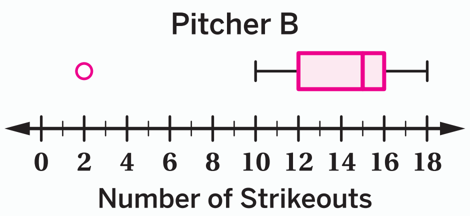
c. Player A. *Explanations vary*. Player A generally scored more points because the median of their data was higher.

Lesson 11

a



Outlier: 18



Outlier: 2

b The mean would be more affected. *Explanations vary*. Every value is used when calculating the mean, so changing any value affects the mean. The median is not affected by the change because 15 is still the middle value after removing the outlier.

Lesson 12

a *Responses and explanations vary*. I would use the median and IQR to compare the rents because the data is not symmetrical. There is an outlier at about $850 which affects the mean and standard deviation.

b *Responses vary*. The median rents for both regions are alike. The rents in the South are more consistent since the rents in southern states have a smaller IQR compared to the IQR of the midwestern states.

Lesson 13

a 0.233

b There is a weak positive association between games won in 2019 and payroll in millions of dollars.

Lesson 14

a There is a weak positive relationship between percentage of tree cover and income in Detroit, Michigan.

b *Responses vary*. Since tree cover and income are not inherently related, someone might show the data to a city council to justify adding more trees into neighborhoods with lower incomes, which would bring the correlation coefficient closer to 0.

Lesson 15

- **a** -2,234: The sale price of a car drops $2,234 every time the age of the car increases by 1 year. 26,250: If a car is new (0 years old), it would sell for $26,250.
- **b** $19,548, or about $20,000

  One strategy for predicting using a model is to substitute the known value into the line of fit.

  $y = -2234(3) + 26250$

  $y = -6702 + 26250$

  $y = 19548$

Lesson 16

- **a** *Responses vary.* The point on the residual plot that represents 8 avocados is positive because the value of the actual data point is greater than the line of fit predicts. The point that represents 3 avocados is negative because the value of the actual data point is less than the line predicts.
- **b** *Responses vary.* The line is not a good fit for the data because many of the points are far from the x-axis. The residuals start off all negative and then turn positive, which shows that the line does not follow the pattern of the data. If the line was a good fit, the residuals would be close to the x-axis and not follow any pattern.

Lesson 17

- **a** Line of Best Fit: $y = -2270.38x + 26886.7$; Correlation Coefficient: $r = -0.96$
- **b** There is a strong negative relationship between car age and sale price.

Lesson 18

- **a** The 0.87 says that there is a strong positive relationship between the number of ice cream cones sold and the number of customers wearing sunglasses.
- **b** *Responses vary.* I don't think one variable causes the other. It doesn't make sense that the number of sunglasses worn or the number of ice cream cones sold cause each other. Both of these likely increase when the weather is warm.

Lesson 19

- **a** *Responses vary.* The number of miles of highway in a city and the amount of time stuck in traffic.
- **b** *Responses vary.* Does the amount of time stuck in traffic increase or decrease as the number of miles of highway increases?
- **c** *Responses vary.* What should we do to help reduce traffic in our community?

Algebra 1 Unit 3 Glossary/Álgebra 1 Unidad 3 Glosario

English	Español

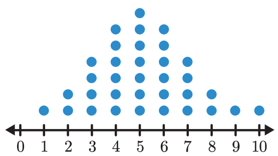
A

association If two variables are statistically related to one another, we say that there is an association between the two variables. A correlation is a type of association.

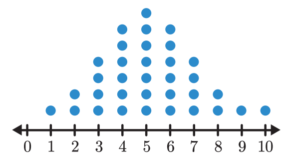
asociación Si dos variables están relacionadas estadísticamente, decimos que existe una asociación entre ambas variables. Una correlación es un tipo de asociación.

B

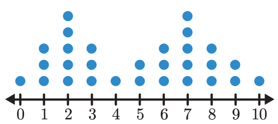
bell-shaped A data distribution whose dot plot or histogram takes the form of a bell with most of the data at the center and fewer points farther from the center.



en forma de campana Una distribución de datos cuyo diagrama de puntos o histograma adopta la forma de una campana donde la mayoría de los puntos están en el centro y hay menos puntos alejados del centro.

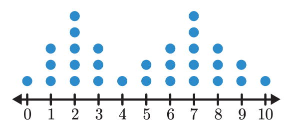


bimodal A data distribution with two very common data values, seen in a dot plot or histogram as two peaks.



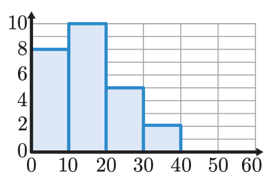
In the dot plot, the two peaks are at 2 and 7.

bimodal Una distribución de datos con dos valores de datos muy comunes y que en un diagrama de puntos o histograma se ve como dos picos.



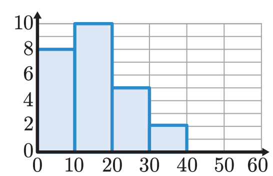
En el diagrama de puntos, los dos picos se encuentran en 2 y 7.

bin (of a histogram) The intervals used to group data values in a histogram are called bins. The distance of each interval is called the bin width.



For example, this histogram shows 4 bins each with a bin width of 10 units.

intervalo (de un histograma) Se denominan intervalos a los espacios que se usan para agrupar los valores de datos en un histograma. La distancia de cada intervalo se denomina ancho del intervalo.



Por ejemplo, este histograma muestra 4 intervalos, cada uno con un ancho de intervalo de 10 unidades.

box plot One way to visualize quantitative data. The data is divided into four sections using five values: the minimum, Q1, Q2 (the median), Q3, and the maximum. A box is drawn between Q1 and Q3, and the line inside the box represents the median.

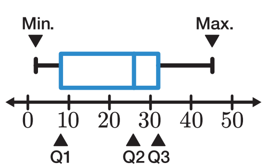
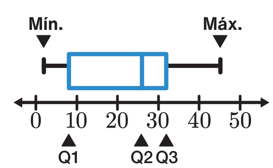


diagrama de caja Una forma de visualizar datos cuantitativos. Los datos se dividen en cuatro secciones utilizando cinco valores: el mínimo, Q1, Q2 (la mediana), Q3 y el máximo. Se dibuja una caja entre Q1 y Q3, y la línea dentro de la caja representa la mediana.



English

C

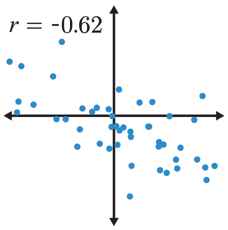
categorical data Data with values that are represented as categories, such as colors, words, or ZIP Codes, instead of values that represent a quantity or measurement.

For example, the breeds of 10 dogs are categorical data. Another example is the colors of 100 flowers.

causation Describes a relationship between two variables in which a change in one variable causes a change in the other variable. A special type of correlation.

correlation A statistical relationship between two or more variables. Also called an association.

correlation coefficient A number between -1 and 1 that describes the strength and direction of a linear association between two numerical variables. The closer the correlation coefficient is to 0, the weaker the linear association. The closer the correlation coefficient is to -1 or 1, the stronger the linear association. It is also called the r-value.



This data set has a correlation coefficient of -0.62.

D

direction (of an association) Describes whether the trend of data on a scatter plot is positive (increasing), negative (decreasing), or neither.

Español

C

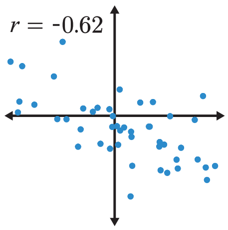
datos categóricos Datos con valores que se representan en categorías, como colores, palabras o códigos postales, en lugar de valores que representan una cantidad o medida.

Por ejemplo, las razas de 10 perros diferentes son datos categóricos. Otro ejemplo son los colores de 100 flores diferentes.

causalidad Describe una relación entre dos variables en la que un cambio en una variable provoca un cambio en la otra variable. Un tipo especial de correlación.

correlación Una relación estadística entre dos o más variables. También se denomina asociación.

coeficiente de correlación Un número entre -1 y 1 que describe la fuerza y la dirección de una asociación lineal entre dos variables numéricas. Cuanto más se acerca el coeficiente de correlación a 0, más débil es la asociación lineal. Cuanto más se acerca el coeficiente de correlación a -1 o 1, más fuerte es la asociación lineal. También se denomina valor r.



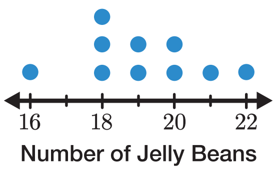
Este conjunto de datos tiene un coeficiente de correlación de -0.62.

D

dirección (de una asociación) Describe si la tendencia de los datos en un diagrama de dispersión es positiva (creciente), negativa (decreciente) o ninguna de las dos.

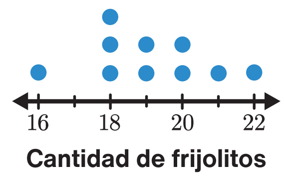
English	Español

dot plot One way to visualize data. Every data point is shown as a dot above its value on a number line. Data sets with repeating values will have multiple dots aligned with the same value.



For example, this dot plot shows that 3 students guessed that there were 18 jelly beans in a jar.

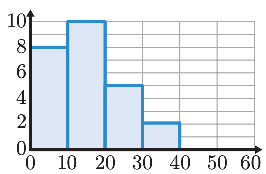
diagrama de puntos Una forma de visualizar datos. Cada punto de datos se muestra como un punto encima de su valor y se apila sobre otros puntos que tienen el mismo valor.



Por ejemplo, esta diagrama de puntos muestra que 3 estudiantes estimaron que había 18 frijolitos de jalea en un tarro.

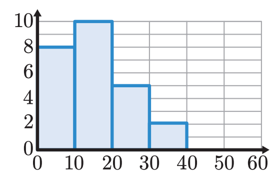
H

histogram One way to visualize quantitative data that groups data into bins, each shown by a rectangle. The height of each rectangle shows how many values are in that bin.



For example, this histogram shows that there are 8 values between 0 and 10.

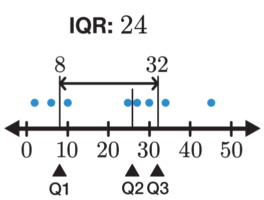
histograma Una forma de visualizar datos cuantitativos que agrupa los datos en intervalos, cada uno de los cuales se representa mediante un rectángulo. La altura de cada rectángulo muestra cuántos valores hay en ese intervalo.



Por ejemplo, este histograma muestra que hay 8 valores entre 0 y 10.

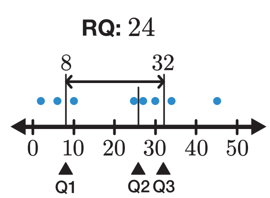
I

interquartile range (IQR) The distance from Q1 to Q3, and the width of the box in a box plot. IQR is a measure of spread.



For example, the IQR of this data set is $32 - 8 = 24$.

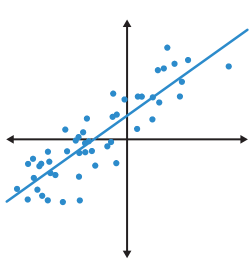
rango intercuartílico (RQ) La distancia de Q1 a Q3, y el ancho de la caja en un diagrama de caja. El RQ es una medida de dispersión.



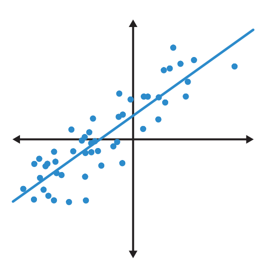
Por ejemplo, el RQ de este conjunto de datos es $32 - 8 = 24$.

L

line of (best) fit The line on a scatter plot that best represents the trend created by the points in a data set. There are many lines of fit for a single data set, but there is only one line of best fit.

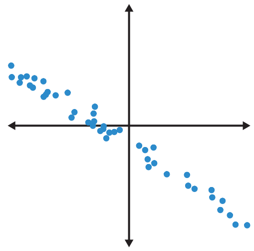


línea de (mejor) ajuste La línea de un diagrama de dispersión que mejor representa la tendencia que definen los puntos de un conjunto de datos. Existen muchas líneas de ajuste para un mismo conjunto de datos, pero solo hay una línea de mejor ajuste.



English

linear association When the points on a scatter plot follow a line, we say there is a linear association between x and y. Linear associations are categorized by their strength and direction.

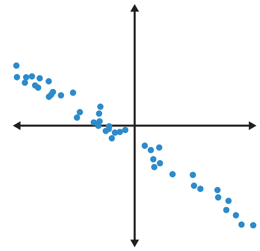


The graph shows a strong, negative linear association.

lurking variable An additional variable that has an effect on the other variables being analyzed. Lurking variables can lead us to make conclusions about data and relationships that are inaccurate.

Español

asociación lineal Cuando los puntos de un diagrama de dispersión siguen una línea, decimos que existe una asociación lineal entre x y y. Las asociaciones lineales se categorizan según su fuerza y dirección.



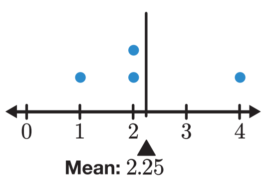
La gráfica muestra una asociación lineal negativa y fuerte.

variable de confusión Una variable adicional que tiene un efecto sobre las demás variables que se analizan. Las variables de confusión pueden producir conclusiones imprecisas en torno a los datos y las relaciones.

M

maximum (of a data set) The greatest value in a data set.

mean A statistic used to describe the typical value of a data set. To calculate the mean, you can add up all the data values and divide by the number of data points. The mean is a measure of center and is also called the average.



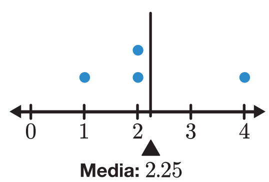
In this data set, the mean is 2.25.

$1 + 2 + 2 + 4 = 9$

$\frac{9}{4} = 2.25$

máximo (de un conjunto de datos) El mayor valor en un conjunto de datos.

media Un dato estadístico que se usa para describir el valor típico de un conjunto de datos. Para calcular la media, se suman todos los valores y el resultado se divide por el número de puntos de datos. La media es una medida de tendencia central y también se denomina promedio.



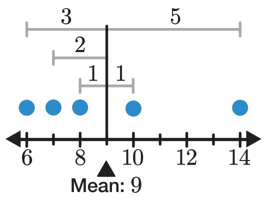
En este conjunto de datos, la media es 2.25.

$1 + 2 + 2 + 4 = 9$

$\frac{9}{4} = 2.25$

Algebra 1 Unit 3 Glossary/Álgebra 1 Unidad 3 Glosario

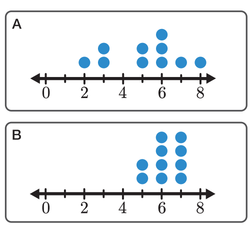
English



mean absolute deviation (MAD) One way to measure how spread out a data set is from the mean. Sometimes we call this the MAD. To calculate the MAD, add the distances between each data value and the mean. Then divide by the number of data values in the set.

In this data set, the MAD is 2.4 because $3+2+1+1+5=12$ and $\frac{12}{5}=2.4$.

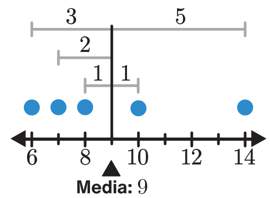
measure of center A single number that summarizes all of the values in a data set. It is usually a typical value for a data set. Mean and median are measures of center.



measure of spread A statistic that measures the variability within a data set is called a measure of spread. The more consistent the data is within a data set, the less variability there is within that data set, and the smaller the measure of spread. Range, interquartile range, and standard deviation are measures of spread.

For example, dot plot A has a larger spread than dot plot B. This means the data represented by dot plot A is less consistent than the data represented by dot plot B.

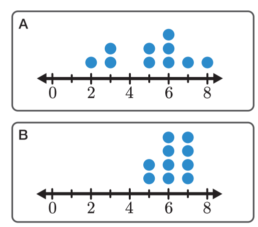
Español



desviación media absoluta (DMA) Una forma de medir cuán disperso está un conjunto de datos con respecto de la media. A veces se denomina DMA. Para calcular la DMA, se suman las distancias entre cada valor y la media. Luego se divide por la cantidad de puntos de datos (valores) del conjunto.

En este conjunto de datos, la DMA es 2.4 porque $3+2+1+1+5=12$ y $\frac{12}{5}=2.4$.

medida de tendencia central, medida de centro Un solo número que resume todos los valores de un conjunto de datos. Suele ser un valor típico de un conjunto de datos. La media y la mediana son medidas de tendencia central.

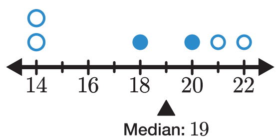


medida de dispersión Se denomina medida de dispersión al dato estadístico que mide la variabilidad de un conjunto de datos. Cuanto más uniformes son los datos en un conjunto, menor es la variabilidad de ese conjunto de datos y menor es la medida de dispersión. El rango, el rango intercuartílico y la desviación estándar son medidas de dispersión.

Por ejemplo, el diagrama de puntos A tiene una dispersión mayor que el diagrama de puntos B. Esto significa que los datos representados en el diagrama de puntos A son menos uniformes que los del diagrama de puntos B.

English

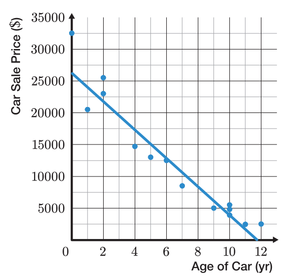
median A statistic used to describe the typical value of a data set. It is the middle value of a data set when the values are in numerical order. If there are two values in the middle of the data set, then the median is the mean of those two values. The median is a measure of center.



minimum (of a data set) The least value in a data set.

N

negative association A relationship between two variables where as one variable increases, the other variable tends to decrease.

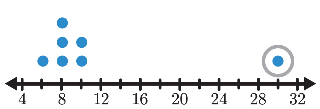


For example, this graph shows a negative association between the age of a car and its sale price. This means that as the age of a car increases, its sale price tends to decrease.

numerical data See *quantitative data*.

O

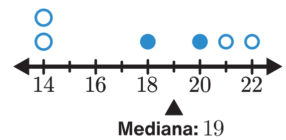
outlier A data value that is far from the other values in the data set. Values that are 1.5 times the IQR from either quartile 1 or quartile 3 are outliers.



The circled point is an outlier.

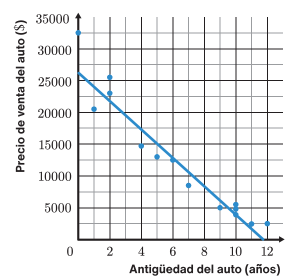
Español

mediana Un dato estadístico que se usa para describir el valor típico de un conjunto de datos. Es el valor del medio de un conjunto de datos cuando los valores se clasifican en orden numérico. Cuando hay dos valores en el medio del conjunto de datos, la mediana es la media de esos dos valores. La mediana es una medida de tendencia central.



mínimo (de un conjunto de datos) El menor valor en un conjunto de datos.

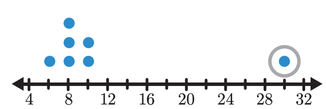
asociación negativa Una relación entre dos variables en la que, a medida que una variable crece, la otra tiende a decrecer.



Por ejemplo, esta gráfica muestra una asociación negativa entre la antigüedad de un auto y su precio de venta. Esto significa que, a medida que aumenta la antigüedad de un auto, su precio de venta tiende a disminuir.

datos numéricos Véase *datos cuantitativos*.

valor atípico Un valor de datos que está lejos de los demás valores del conjunto de datos. Son valores atípicos aquellos que son 1.5 veces el RQ del cuartil 1 o del cuartil 3.



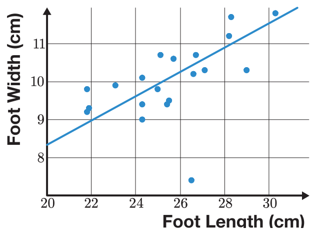
El punto encerrado en un círculo es un valor atípico.

Algebra 1 Unit 3 Glossary/Álgebra 1 Unidad 3 Glosario

P

English

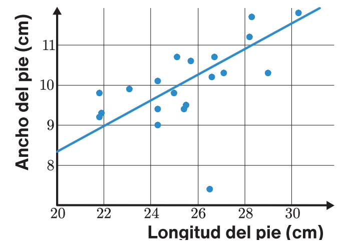
positive association A relationship between two variables where as one variable increases, the other variable tends to increase.



For example, this graph shows a positive association between foot length and width. This means that as foot length increases, foot width also increases.

Español

asociación positiva Una relación entre dos variables en la que, a medida que una variable crece, la otra tiende a crecer.



Por ejemplo, esta gráfica muestra una asociación positiva entre la longitud y el ancho del pie. Esto significa que, a medida que aumenta la longitud del pie, también aumenta su ancho.

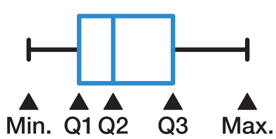
Q

English

quantitative data Data represented as numbers, quantities, or measurements that can be meaningfully compared is called quantitative data, or numerical data.

How many pets do you have? is a question that produces quantitative data. This is different from categorical data.

quartile Quartiles divide a data set into four sections.
Quartile 1 is the median of the lower half of the data.
Q2 is the median.
Q3 is the median of the upper half of the data.
Q4 is the maximum.

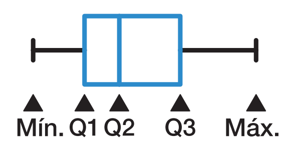


Español

datos cuantitativos Los datos que se representan como números, cantidades o medidas y que pueden compararse de forma significativa se denominan datos cuantitativos o datos numéricos.

¿Cuántas mascotas tienes? es una pregunta que produce datos cuantitativos, que son diferentes de los datos categóricos.

cuartil Los cuartiles dividen un conjunto de datos en cuatro secciones.
El cuartil 1 es la mediana de la mitad inferior de los datos.
Q2 es la mediana.
Q3 es la mediana de la mitad superior de los datos.
Q4 es el máximo.



R

English

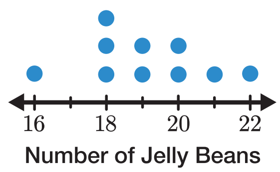
r-value See *correlation coefficient*.

Español

valor r Véase *coeficiente de correlación*.

English

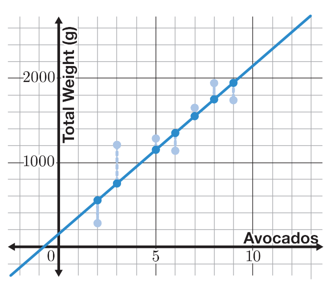
range (of a data set) The difference between the maximum and minimum values in a data set. Range is a measure of spread.



For example, the range of this data set is 6 jelly beans because $22 - 16 = 6$.

relative frequency table A type of two-way table used to compare data across two categorical variables. Relative frequency tables present the fraction or percent of the data that is in that category, instead of the actual number of data points. You can use this representation to see if the data presents evidence that there is an association between the two variables.

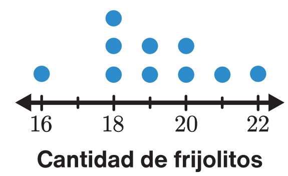
residual The difference between the y-value for a point in a scatter plot and the value predicted by the line of best fit.



The dashed lines on the graph show the residuals for each data point.

Español

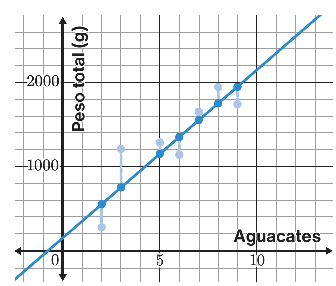
rango (de un conjunto de datos) La diferencia entre los valores máximo y mínimo de un conjunto de datos. El rango es una medida de dispersión.



Por ejemplo, el rango de este conjunto de datos es 6 frijolitos de jalea porque $22 - 16 = 6$.

tabla de frecuencia relativa Un tipo de tabla de doble entrada que se usa para comparar datos entre dos variables categóricas. Las tablas de frecuencia relativa presentan la fracción o el porcentaje de los datos que están en esa categoría, en lugar de la cantidad de puntos de datos. Esta representación puede emplearse para saber si los datos muestran que existe una asociación entre las dos variables.

residuo La diferencia entre el valor y de un punto en un diagrama de dispersión y el valor que predice la línea de mejor ajuste.

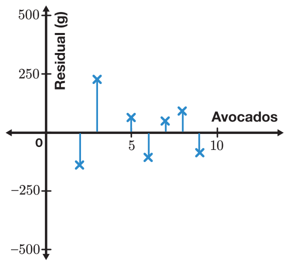


Las líneas discontinuas en la gráfica muestran los residuos de cada punto de datos.

Algebra 1 Unit 3 Glossary/Álgebra 1 Unidad 3 Glosario

English

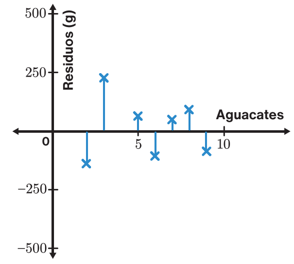
residual plot A scatter plot of residual values for a data set. The x-axis represents the values predicted by the line of best fit, and the y-value of each point represents the value of the residual.



The residual plot shows how the total weight of different numbers of avocados vary from their predicted values.

Español

diagrama de residuos Un diagrama de dispersión de los valores residuales de un conjunto de datos. El eje x representa el valor que predice la línea de mejor ajuste, y el valor y de cada punto representa el valor del residuo.



El diagrama de residuos muestra cómo el peso total de diferentes cantidades de aguacates varía con respecto a sus valores esperados.

S

scatter plot A graph of plotted points that shows the relationship between two variables.

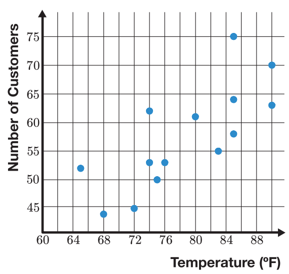
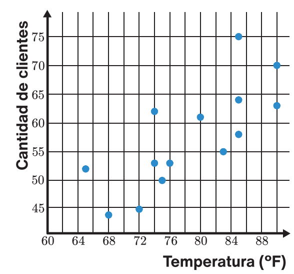


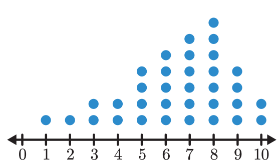
diagrama de dispersión Un diagrama en el que los puntos graficados muestran la relación entre dos variables.



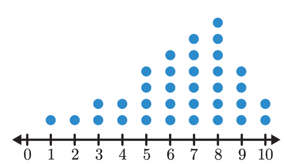
shape (of a data set) The description of the distribution (or pattern) of the data within a data set. We often describe the shape of a data set with terms like bell-shaped, bimodal, clustered, uniform, symmetric, skewed, and outlier.

forma (de un conjunto de datos) La descripción de la distribución (o patrón) de los datos en un conjunto de datos. A menudo, la forma de un conjunto de datos se describe con términos como en forma de campana, bimodal, agrupada, uniforme, simétrica, asimétrica y con valores atípicos.

skewed A data distribution where there are more values concentrated on one end of the data, and few values on the other end.

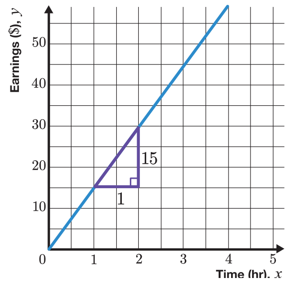


asimétrico Una distribución de datos en la que hay más valores concentrados en un extremo de los datos y pocos valores en el otro extremo.



English

slope A number that describes the direction and steepness of a line. Slope represents the amount that y changes when x increases by 1. Since the slope between any two points on a line will be the same, we can say that a line has a constant rate of change. One way to calculate slope is to divide the vertical distance between any two points on the line by the horizontal distance between those points.

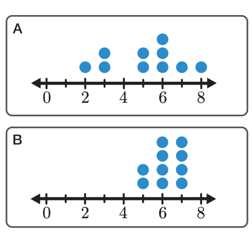


In this graph, y increases by 15 dollars when x increases by 1 hour. The slope of the line is 15, and the rate of change is 15 dollars per hour.

slope-intercept form A way to write a linear equation that highlights the slope and the y-intercept of the line it represents. Slope-intercept form equations are written as $y = mx + b$, where m represents the slope, b represents the y-intercept of the line, and x and y are variables.

The equations $y = 2x + 4$ and $y = -5x - 10$ are in slope-intercept form. The equation $2x + 5y = 20$ is not in slope-intercept form.

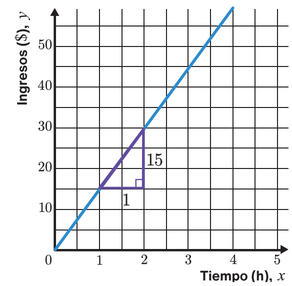
spread A description of how the data values in the distribution vary from the center of the distribution. The spread also describes the variability of a distribution. A measure of spread is a single value that can be used to summarize the consistency of the values in a data set, such as range, MAD, or IQR.



For example, dot plot A has a larger spread than dot plot B.

Español

pendiente Un número que describe la dirección e inclinación de una recta. La pendiente representa la cantidad en la que cambia y cuando x se incrementa en 1. Como la pendiente entre dos puntos cualesquiera de una recta es la misma, podemos decir que una recta tiene una tasa de cambio constante. Una manera de calcular la pendiente es dividir la distancia vertical entre dos puntos cualesquiera en la recta por la distancia horizontal entre dichos puntos.

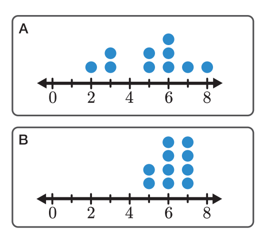


En esta gráfica, y incrementa en 15 dólares cuando x incrementa en 1 hora. La pendiente de la recta es 15 y la tasa de cambio es de 15 dólares por hora.

forma pendiente-intersección Una forma de escribir una ecuación lineal que destaca la pendiente y la intersección con el eje y de la recta que representa. Las ecuaciones en forma pendiente-intersección se escriben como $y = mx + b$, donde m representa la pendiente, b representa la intersección con el eje y de la recta, y tanto x como y son variables.

Las ecuaciones $y = 2x + 4$ y $y = -5x - 10$ están en forma pendiente-intersección. La ecuación $2x + 5y = 20$ no está en forma pendiente-intersección.

dispersión Una descripción de cómo varían los valores de los datos en la distribución desde el centro de la distribución. La dispersión también describe la variabilidad de una distribución. Una medida de dispersión es un valor único que puede emplearse para resumir la uniformidad de los valores de un conjunto de datos; por ejemplo, DMA o RQ.

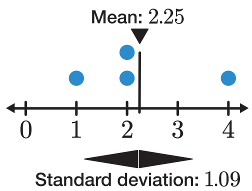


Por ejemplo, el diagrama de puntos A tiene una dispersión mayor que el diagrama de puntos B.

Algebra 1 Unit 3 Glossary/Álgebra 1 Unidad 3 Glosario

English

standard deviation A statistic that measures how spread out the values in a data set are from its mean. The larger the standard deviation, the further the data points within a data set are from its mean. Standard deviation is calculated by a method similar to calculating the MAD (mean absolute deviation). The exact method is studied in more advanced courses.



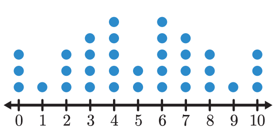
This dot plot shows data with a mean of 2.25 and a standard deviation of 1.09.

statistic A quantity that is calculated from a data set, such as the mean, median, standard deviation, or IQR.

strength (of an association) Describes how closely the data on a scatter plot match a statistical model. The strength of an association can be described as strong (well-predicted by the model) or weak (not well-predicted by the model).

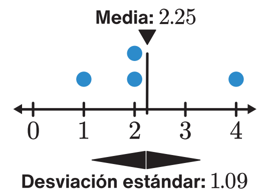
survey A method of collecting data by asking people to answer a given set of questions.

symmetric A data distribution that has a vertical line of symmetry.



Español

desviación estándar Un dato estadístico que mide la dispersión de los valores de un conjunto de datos con respecto a su media. Cuanto mayor es la desviación estándar, más se alejan los puntos de datos de un conjunto de su media. La desviación estándar se calcula mediante un método similar al cálculo de la DMA (desviación media absoluta). El método exacto se estudia en cursos más avanzados.



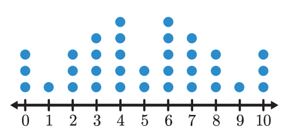
Este diagrama de puntos muestra datos con una media de 2.25 y una desviación estándar de 1.09.

dato estadístico Una cantidad que se calcula a partir de un conjunto de datos, como la media, la mediana, la desviación estándar o el RQ.

fuerza (de una asociación) Describe la medida en que los datos de un diagrama de dispersión coinciden con un modelo estadístico. La fuerza de una asociación puede describirse como fuerte (el modelo la predice bien) o débil (el modelo no la predice bien).

encuesta Un método de recopilación de datos que consiste en pedir a los encuestados que respondan a una serie de preguntas.

simétrico Una distribución de datos que tiene un eje vertical de simetría.



English

T

total relative frequency table A type of two-way table used to compare data across two categorical variables. Total relative frequency tables present the fraction or percentage of the data that is in that category, instead of the number of data points. You can use this representation to see if the data presents evidence that there is an association between the two variables.

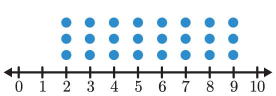
	Meditated	Did Not Meditate	Total
Calm	46%	8%	54%
Agitated	24%	22%	46%
Total	70%	30%	100%

two-way table A way to compare two categorical variables. It shows one of the variables across the top and the other down one side. Each entry in the table is the frequency or relative frequency of the category shown by the column and row headings.

	Meditated	Did Not Meditate	Total
Calm	45	8	53
Agitated	23	21	44
Total	68	29	97

U

uniform A data distribution that has the data values evenly distributed.



Español

T

tabla de frecuencia relativa total Un tipo de tabla de doble entrada que se usa para comparar datos entre dos variables categóricas. Las tablas de frecuencia relativa total presentan la fracción o el porcentaje de los datos que se encuentran en esa categoría, en lugar de la cantidad de puntos de datos. Esta representación puede emplearse para saber si los datos muestran que existe una asociación entre las dos variables.

	Meditaron	No meditaron	Total
Calmados	46%	8%	54%
Agitados	24%	22%	46%
Total	70%	30%	100%

tabla de doble entrada Una manera de comparar dos variables categóricas. Muestra una de las variables en la fila superior y la otra variable en un costado. Cada entrada de la tabla es la frecuencia o frecuencia relativa de la categoría que describen los títulos de la columna y fila.

	Meditaron	No meditaron	Total
Calmados	45	8	53
Agitados	23	21	44
Total	68	29	97

U

uniforme Una distribución de datos que tiene los valores de los datos distribuidos uniformemente.