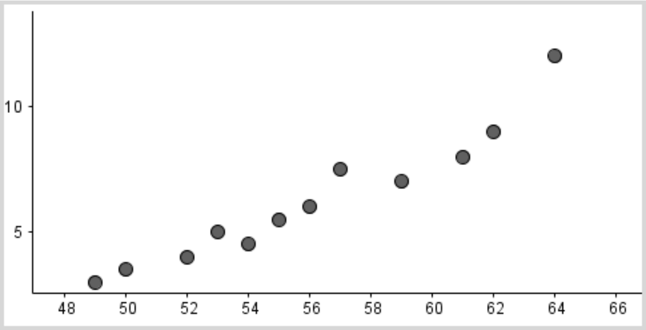
SCATTER PLOTS AND INTERPRETING GRAPHS



Unit Overview

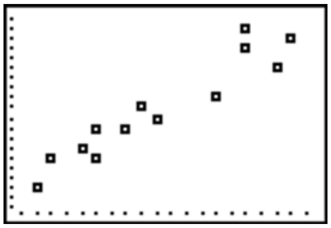
Scatter plots are used to illustrate how two variables relate to each other by showing how closely the data points cluster to a line (line of best fit). Scatter plots can be used to predict relationships between two sets of data, such as those related to weather. Lastly, you will also learn about 2-way frequency tables.

Scatter Plots

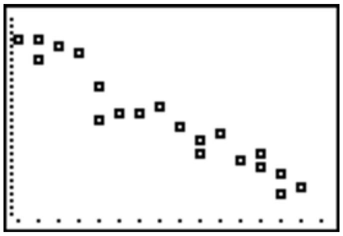
Scatter plots are an easy way to determine if there is a relationship between two variables. This relationship is called a **correlation**. A correlation is based on the slope of the line of best fit. (We will discuss how to find the line of best fit later in the unit).

There are three possible types of correlation: a) positive, b) negative, or c) no correlation. The illustrations below show the graph of each correlation.

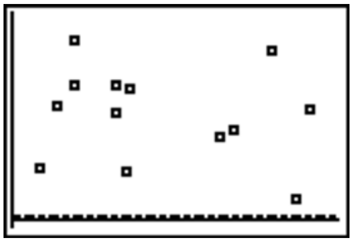
a.)



b.)



c.)



In graph "a," notice how the points cluster in a rise to the right. Recall from a previous unit that this suggests a **positive slope**. Graph "b" shows points that cluster in a fall to the right, which suggests a **negative slope,** and graph "c" shows no cluster pattern and suggests that the two variables have **no relationship** to each other.

Let's take a look at a few examples and determine if each situation has a positive, negative, or no correlation.

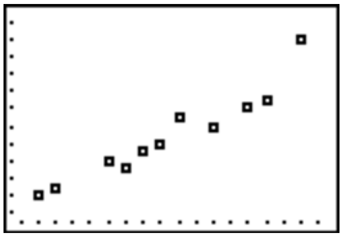
Example: Determine which scatter plot represents each situation.

a.) a person's height and his/her hourly wage

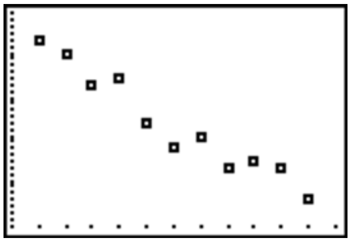
b.) a person's height and his/her shoe size

c.) a person's age and his/her time needed to run 100 yards

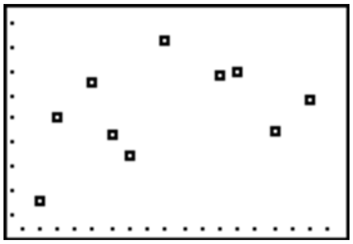
1)



2)



3)



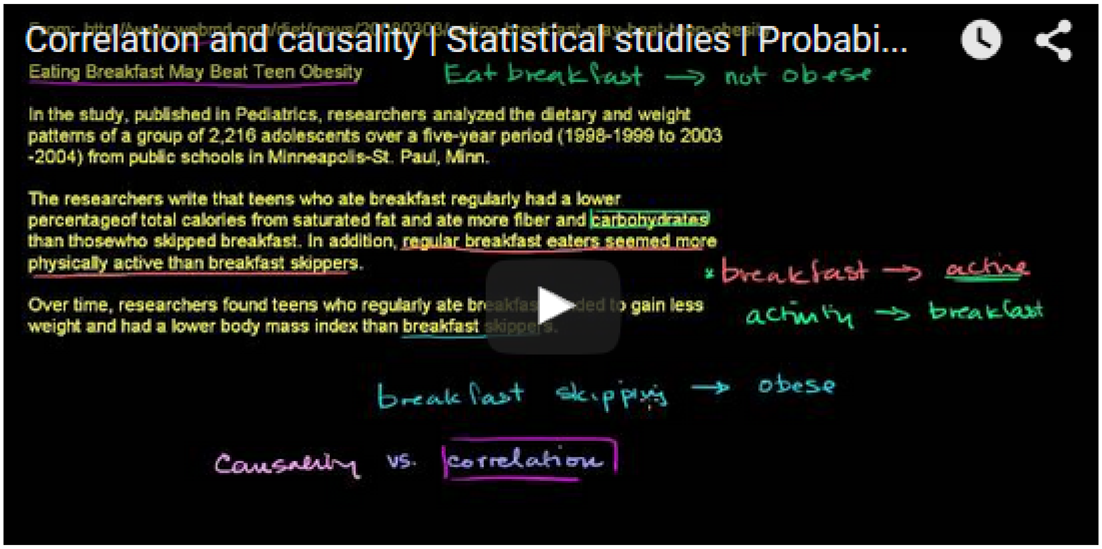
Scatter plot 1 shows a strong positive correlation. A **positive correlation** occurs when **both variables increase**. As you grow taller your shoe size increases; therefore, **plot 1** represents situation **"b."**

Scatter plot 2 shows a strong negative correlation. A **negative correlation** occurs when **one variable increases as the other variable decreases**. In situation **"c"** as your age increases, the time it takes you to run 100 yards decreases.

The third scatter plot shows **no correlation** because the data points are **randomly scattered**. Your height has no relationship with your hourly wage; therefore, this plot represents situation **"a."**

*Watch this video to learn more about correlation and causality:

Click on the link to watch the video "Correlation and causality" or click on the video.



Correlation and causality | Statistical studies | Probabi...

Eating Breakfast May Beat Teen Obesity

In the study, published in Pediatrics, researchers analyzed the dietary and weight patterns of a group of 2,216 adolescents over a five-year period (1998-1999 to 2003 -2004) from public schools in Minneapolis-St. Paul, Minn.

The researchers write that teens who ate breakfast regularly had a lower percentageof total calories from saturated fat and ate more fiber and carbohydrates than thosewho skipped breakfast. In addition, regular breakfast eaters seemed more physically active than breakfast skippers.

Over time, researchers found teens who regularly ate breakfast tended to gain less weight and had a lower body mass index than breakfast skippers.

Eat breakfast → not obese
* breakfast → active
activity → breakfast
breakfast skipping → obese
Causality vs. correlation

Stop! **Go to Questions #1-8 about this section, then return to continue on to the next section.**

Line of Best Fit

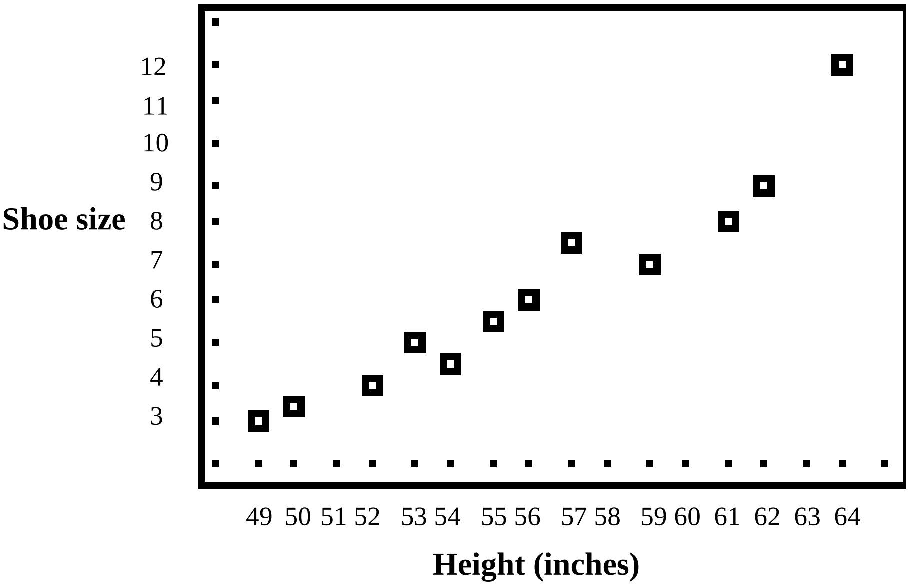
Earlier in this unit, we talked about a line of best fit. Data in a scatter plot can be studied using a line of best fit which represents the trend or behavior of the data. A line of best fit can be used to predict what the data might be for values not given.

Let's use the data given from the example representing height in inches and shoe size to find the line of best fit for the scatter plot.

Example:

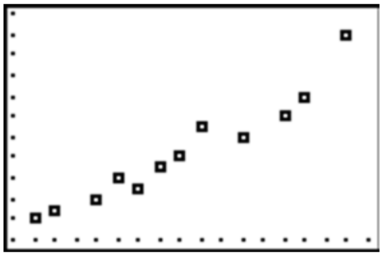
Height **(inches)**	49	50	52	53	54	55	56	57	59	61	62	64
Shoe size	3	3.5	4	5	4.5	5.5	6	7.5	7	8	9	12

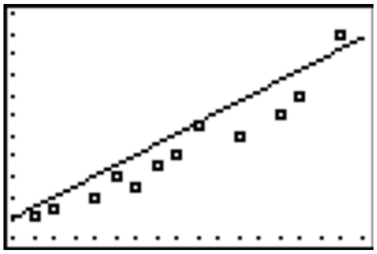
1) Plot the data as coordinates on a coordinate plane (height, shoe size). The height on the horizontal or *x*-axis and the shoe size on the vertical or *y*-axis.

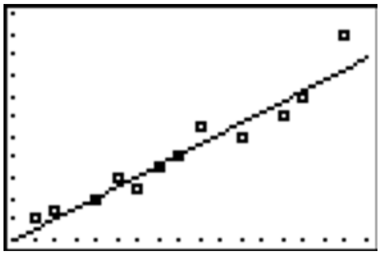


2) Use some type of a straight edge, such as a clear ruler or piece of uncooked spaghetti, to model the line that represents the trend of the data.

3) To fit the line to the points, place your straightedge so that it best matches the overall trend-having the same number of points above and below the line. The examples below show what not to do and what should be done. ***The line does not have to go through any of the points as long as the general pattern or trend is represented.**


Scatter Plot


Line 1


Line 2

Line 1 goes through one of the points but completely ignores the others.

Line 2 is a closer representation of the data points as you can see it takes all points into consideration and there seems to be as many points above the line as there are below the line. Therefore, Line 2 is a better fit.

It will take you a while to practice this and it is a concept that has no exact answer. The idea of a line of best fit is to predict what the data will show for points that are not plotted.

After you have established the trend of the data, you need to choose two points that lie on your line (x_1, y_1) and (x_2, y_2), so you can calculate the equation for the line of best fit.

Recall from a previous unit that if given two points on a line, you can find the equation of the line. Follow the directions below to refresh your memory.

1) Find the slope using the coordinates of two points, but do not choose two points that are side by side. *It's a good idea to use a point from the first part of the line and the last part of the line.*

$$m = \frac{y_2 - y_1}{x_2 - x_1}$$

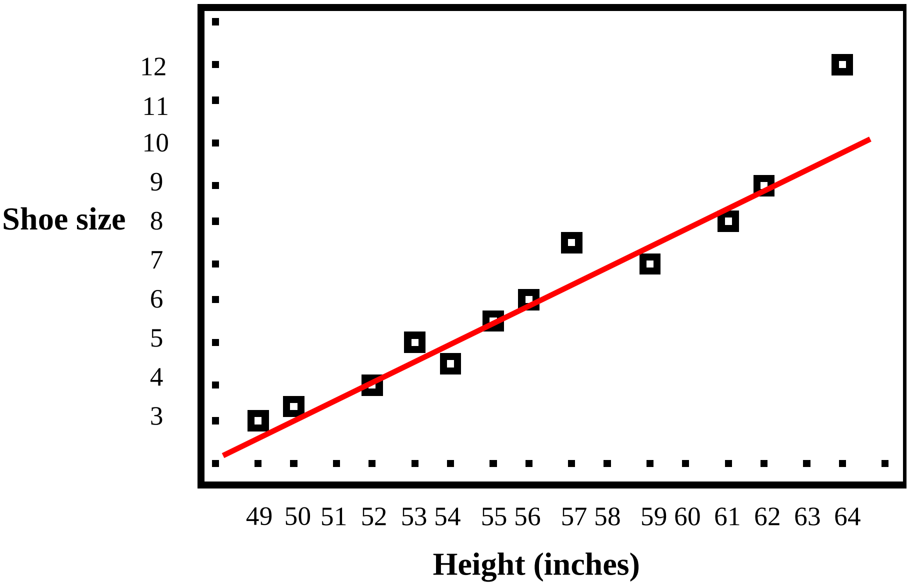
2) Use the coordinates of one of the points and the slope that you just found in the point-slope formula to generate the equation.

$$y - y_1 = m(x - x_1)$$

3) Solve the equation for *y* and you have the equation of the line of best fit.

4) We will choose two points that look like they lie on the line and are from the first part of the line and the last part of the line.

Let's choose (52, 4) and (62, 9).



5) Find the slope using the coordinates of the two points.

$$m = \frac{y_2 - y_1}{x_2 - x_1} = \frac{9-4}{62-52} = \frac{5}{10} = \frac{1}{2} \qquad x_1 = 52 \quad y_1 = 4 \quad x_2 = 62 \quad y_2 = 9$$

6) Use the coordinates of one of the points and the slope in the point-slope formula. We'll use (52, 4).

$$y - y_1 = m(x - x_1) \qquad x_1 = 52 \quad y_1 = 4 \quad m = \frac{1}{2}$$

$$y - 4 = \frac{1}{2}(x - 52)$$

$$y - 4 = \frac{1}{2}x - 26$$

$$y = \frac{1}{2}x - 22$$

The equation of the line of best fit is $y = \frac{1}{2}x - 22$. You can now use this equation to determine data that is not shown on the graph.

For example, let's say that you wanted to know the shoe size of someone who is 68 inches tall. Since your scatter plot does not represent data for this height, you can use the equation $y = \frac{1}{2}x - 22$ to determine this.

Since you know the height of the person (68 inches), you will replace this value for *x* in the equation and solve for *y*, which represents the shoe size.

$$y = \frac{1}{2}(68) - 22$$

$$y = 34 - 22$$

$$y = 12$$

The trend of the data shows us that if a person is 68 inches tall, they will probably wear a shoe size of 12. Again, this is just a prediction based on the information we have in our scatter plot. It does not mean that a person 68 inches tall should or will wear a size 12 shoe.

Click on the activity below to practice making a scatterplot and calculating a line of best fit.

Scatterplot - NLVM

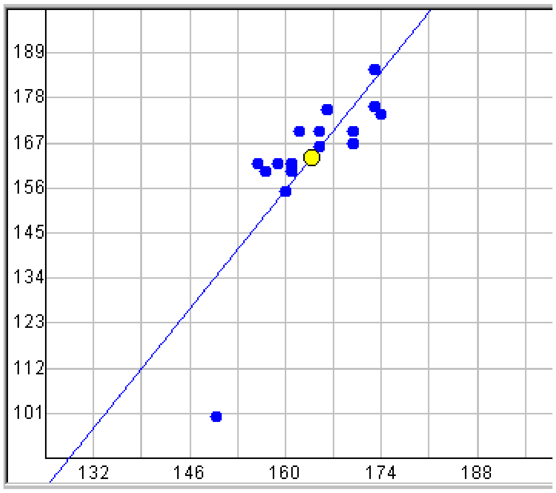
A graphing utility will come up that allows you to enter many data points. Enter the following data for arm span and height.

Arm span	Height
165	166
173	176
173	185
156	162
160	155
170	170
157	160
170	167
162	170
159	162
161	160
161	162

Enter the arm span as *x* and the height as *y*. The points are automatically plotted and you can see the line of best fit is also shown. Underneath, where you see "Scale," you can adjust the axes so you can better see the data. Change the *x* min to 140 and *x* max to 200. Change *y* min to 150 and *y* max to 190. Then click "Apply." This makes the data more in the center of the window for you to see. Notice how the line of best fit is fitted to be through the middle of most points. It is calculated so that the distance between the points and the line is minimal.

You can see the equation for the line of best fit in the section to the right of the "Scale" box. You will see several things like *n* = #, *r* = #, etc. You will also see *y* = something. This should be your line of best fit. Look at the created graph and pick two points you feel would give you the line of best fit and see how closely it matches the equation given on screen. Were you close? The slopes should be fairly similar.

Look at this scatterplot.

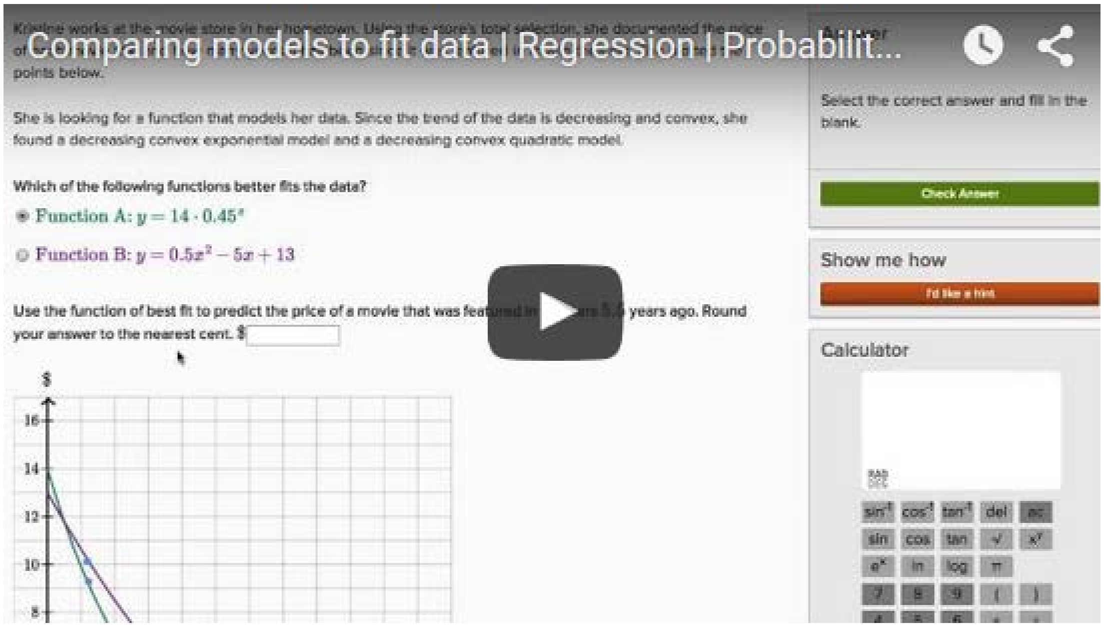


Notice that there is one point, lonely down towards the bottom. This point is (150, 100). We call this an outlier. This is a point that doesn’t seem to fit with the others. When you have outliers on a scatterplot, you will want to make your best fit line go through the majority of points, so do not consider the outliers. What about your equation of your best fit line? Use two points that fit the line. When looking at data, it’s important to understand overall trend and which points do not seem to fit the trend.

Scatterplots do not always have to follow a linear path. **You may get points in an exponential or quadratic path as well.**

Watch this video for an example where this may happen:

Click on the link to watch the video "Comparing models to fit data" or click on the video.



Comparing models to fit data | Regression | Probabilit...

Kristine works at the movie store in her hometown. Using the store's total selection, she documented the price of [illegible] points below.

She is looking for a function that models her data. Since the trend of the data is decreasing and convex, she found a decreasing convex exponential model and a decreasing convex quadratic model.

Which of the following functions better fits the data?

● Function A: $y = 14 \cdot 0.45^x$

○ Function B: $y = 0.5x^2 - 5x + 13$

Use the function of best fit to predict the price of a movie that was featured in [illegible] years ago. Round your answer to the nearest cent. $ □

Answer

Select the correct answer and fill in the blank.

Practice answering questions about these types of scatterplots: Fitting quadratic and exponential functions to scatter plots.

Stop! **Go to Questions #9-15 about this section, then return to continue on to the next section.**

2-Way Frequency Tables

Two-way frequency tables show frequency in comparison to two variables. An example of this is shown below.

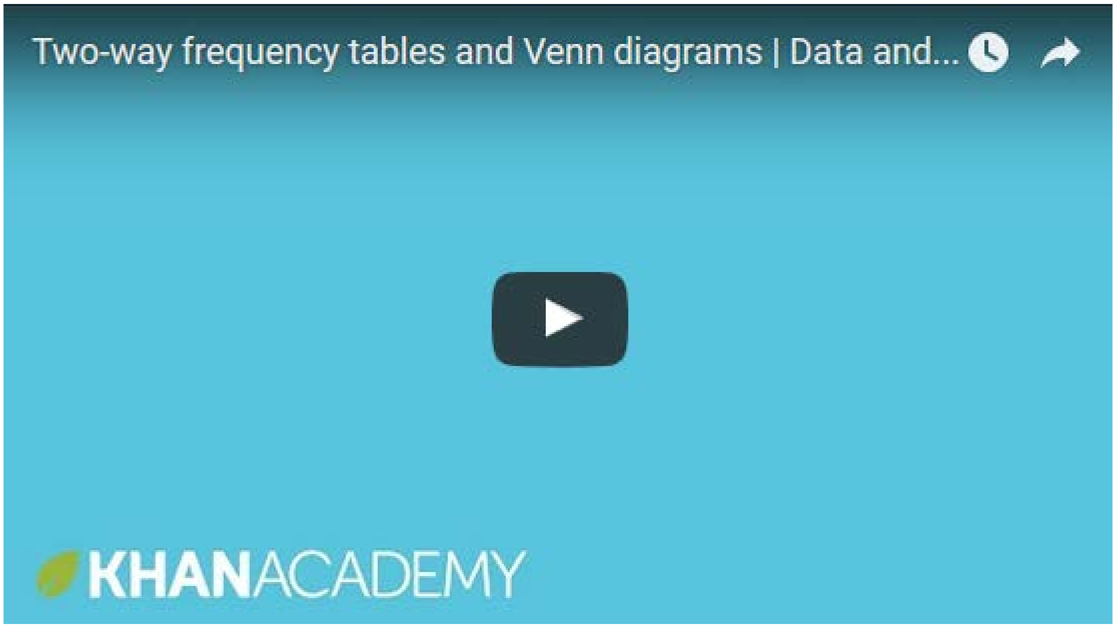
Sara has surveyed her class to see if there is a correlation between gender and in which course which they do well. She asked them if they scored better in English or in Math. Here are the results.

	Male	Female	Total
English	15	20	35
Math	25	15	40
Total	40	35	75

From this table, we can find the number of males who score better in English, the number of females who score better in English, or the total number of students that score better in English. We can also read this table for the number of females or to find the same answers for Math.

*Watch this video on creating 2-Way Frequency Tables. Please watch the introduction so you know what the data is. Then you may fast forward (as we are not discussing Venn Diagrams at this time) to time 3:36 on the video. Then, watch to the end.

Click on the link to watch the video "Two-way frequency tables and Venn diagrams" or click on the video.



Example #1: Make a 2-Way Frequency Table for the following data:

Susan surveyed her class and found the following: 8 girls have blue eyes and 15 have brown eyes. She also found that 4 boys have blue eyes and 12 have brown eyes.

First, we set up the table. You can put either categories along the top or along the bottom. Let's use gender for the rows and eye color for the columns.

	Blue	Brown	Total
Male			
Female			
Total			

Now, fill in what you know. Then, add the columns and rows to fill in the totals.

	Blue	Brown	Total
Male	4	12	16
Female	8	15	23
Total	12	27	39

Notice that both total columns should add to the same amount in the bottom right-hand box. If not, there is a mistake somewhere.

What are the trends here? Do males or females tend to have a particular color eye? Or is a particular color of eye more dominant? It seems that for both male and females, more have brown eyes and gender does not seem to matter.

Relative Frequencies

Sometimes, we will use relative frequencies when making a 2-way chart. A relative frequency is found by finding the decimal equivalent of (# in category) / (total number). For example, in our chart above, the relative frequency for females with blue eyes would be 8/39. Rounding to 3 decimal places would give us 0.205. Since we are giving the values as decimals, all values should then add to 1.000 for our total in the bottom right hand box.

Try this! Fill in the table with the relative frequencies. Use the data above.

"Click here" to check your answer.

	Blue	Brown	Total
Male	.103	.308	.410
Female	.205	.385	.590
Total	.308	.692	1.000

Remember that there were 39 total participants in the survey, so all numbers are divided by 39.

To practice understanding two-way frequency tables, please go to: Two-way relative frequency tables

For practice interpreting two-way tables, please go to: Interpreting two-way tables

Stop! **Go to Questions #16-30 to complete this unit.**