

## 8.7: Correlation and Causation, Scatter Plots

The label on a can of Planters Cocktail Peanuts says, “Scientific evidence suggest but does not prove that eating 1.5 ounces per day of most nuts, such as peanuts, as part of a diet low in saturated fat and cholesterol & not resulting in increased caloric intake may reduce the risk of heart disease. See nutritional information for fat content (1.5 oz. is about 53 pieces).” Why is it written this way and what does this statement mean? There are many studies that exist that show that two variables are related to one another. The strength of a relationship between two variables is called **correlation**. Variables that are strongly related to each other have strong correlation. However, if two variables are correlated it does not mean that one variable caused the other variable to occur. The above example from the Planters Cocktail Peanuts label is an example of this. There is a strong correlation between eating a diet that is low in saturated fat and cholesterol and heart disease. But that correlation does not mean that eating a diet that is low in saturated fat and cholesterol will cause your risk of heart disease to go down. There could be many different variables that could cause both variables in question to go down or up. One example is that a person’s genetic makeup could make them not want to eat fatty food and also not develop heart disease. No matter how strong a correlation is between two variables, you can never know for sure if one variable causes the other variable to occur without conducting experimentation. The only way to find out if eating a diet low in saturated fat and cholesterol actually lowers the risk of heart disease is to do an experiment. This is where you tell one group of people that they have to eat a diet low in saturated fat and cholesterol and another group of people that they have to eat a diet high in saturated fat and cholesterol, and then observe what happens to both groups over the years. You cannot morally do this experiment, so there is no way to prove the statement. That is why the word “may” is in the statement. We see many correlations like this one. Always be sure not to make a correlation statement into a causation statement.

### Example 8.7.1: Correlation vs Causation

For each of the following scenarios answer the question and give an example of another variable that could explain the correlation.

1. There is a negative correlation between number of children a woman has and her life expectancy. Does that mean that having children causes a woman to die earlier?

A correlation between two variables does not mean that one causes the other. A possible cause for both variables could be better health care. If there is better health care, then life expectancy goes up, and also with better health care birth control is more readily available.

2. There is a positive correlation between ice cream sales and the number of drownings at the beach. Does that mean that eating ice cream can cause a person to drown?

A correlation between two variables does not mean that one causes the other. The cause for both could be that the temperature is going up. The higher the temperature, the more likely someone will buy ice cream and the more people at the beach.

3. There is a correlation between waist measures and wrist measures. Does this mean that your waist measurement causes your wrist measurement to change?

A correlation between two variables does not mean that one causes the other. The cause of both could be a person’s genetics, eating habits, exercise habits, etc.

How do we tell if there is a correlation between two variables? The easiest way is to graph the two variables together as ordered pairs on a graph called a **scatter plot**. To create a scatter plot, consider that one variable is the independent variable and the other is the dependent variable. This means that the dependent variable depends on the independent variable. We usually set up these two variables as ordered pairs where the independent variable is first and the dependent variable is second. Thus, when graphed, the independent variable is graphed along the horizontal axis and the dependent variable is graphed along the vertical axis. You do not connect the dots after plotting these ordered pairs. Instead look to see if there is a pattern, such as a line, that fits the data well. Here are some examples of scatter plots and how strong the linear correlation is between the two variables.

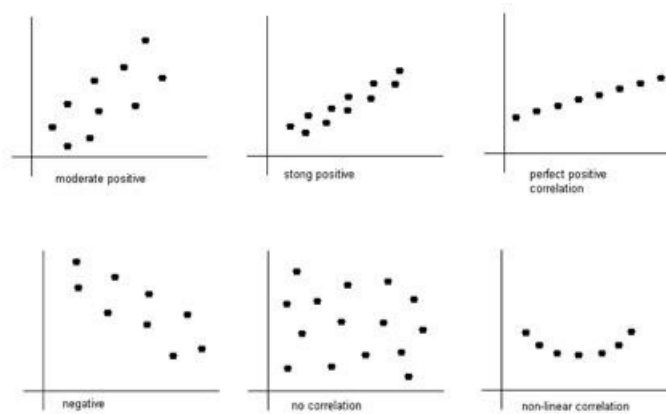


Figure 8.7.1: Scatter Plots Showing Types of Linear Correlation

Creating a scatter plot is not difficult. Just make sure that you set up your axes with scaling before you start to plot the ordered pairs.

### Example 8.7.2: Creating a Scatter Plot

Data has been collected on the life expectancy and the fertility rate in different countries ("World health rankings," 2013). A random sample of 10 countries was taken, and the data is:

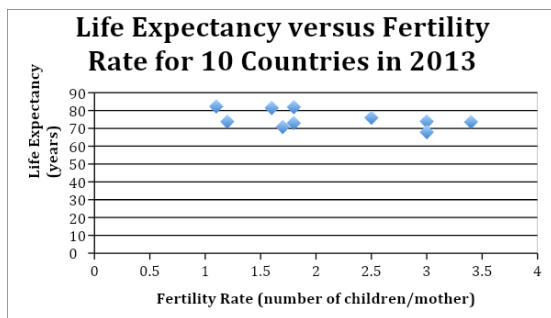
Table 8.7.2: Life Expectancy and Fertility Rate in 2013

Country	Life Expectancy (years)	Fertility Rate (number of children per mother)
SINGAPORE	82.3	1.1
MONACO	81.9	1.8
CANADA	81.5	1.6
ECUADOR	76	2.5
MALAYSIA	73.9	3
LITHUANIA	73.8	1.2
BELIZE	73.6	3.4
ALGERIA	73	1.8
TRINIDAD/TOB.	70.8	1.7
TAJIKISTAN	67.9	3

To make the scatter plot, you have to decide which variable is the independent variable and which one is the dependent variable. Sometimes it is obvious which variable is which, and in some case it does not seem to be obvious. In this case, it seems to make more sense to predict what the life expectancy is doing based on fertility rate, so choose life expectancy to be the dependent variable and fertility rate to be the independent variable. The horizontal axis needs to encompass 1.1 to 3.4, so have it range from zero to four, with tick marks every one unit. The vertical axis needs to encompass the numbers 70.8 to 81.9, so have it range from zero to 90, and have tick marks every 10 units.

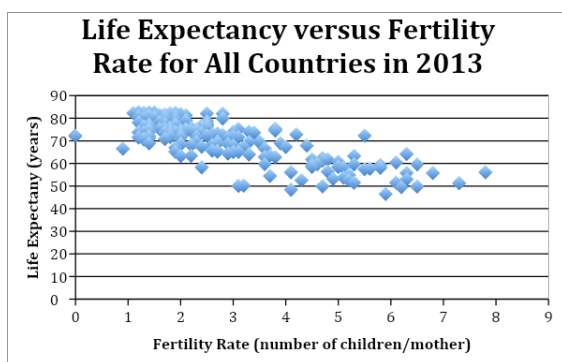
*Note: Always start the vertical axis at zero to avoid exaggeration of the data.*

### Graph 8.7.3: Scatter Plot of Life Expectancy versus Fertility Rate



From the graph, you can see that there is somewhat of a downward trend, but it is not prominent. What this says is that as fertility rate increases, life expectancy decreases. The trend is not strong which could be due to not having enough data or this could represent the actual relationship between these two variables. Let's see what the scatter plot looks like with data from all countries in 2013 ("World health rankings," 2013).

**Graph 8.7.4: Scatter Plot of Life Expectancy versus Fertility Rate for All Countries in 2013**



Again, there is a downward trend. It looks a little stronger than the previous scatter plot and the trend looks more obvious. This correlation would probably be considered moderate negative correlation. It appears that there is a trend that the higher the fertility rate, the lower the life expectancy. Caution: just because there is a correlation between higher fertility rate and lower life expectancy, do not assume that having fewer children will mean that a person lives longer. The fertility rate does not necessarily cause the life expectancy to change. There are many other factors that could influence both, such as medical care and education. Remember a correlation does not imply causation.

This page titled [8.7: Correlation and Causation, Scatter Plots](#) is shared under a [CC BY-SA 4.0](#) license and was authored, remixed, and/or curated by [Maxie Inigo, Jennifer Jameson, Kathryn Kozak, Maya Lanzetta, & Kim Sonier](#) via [source content](#) that was edited to the style and standards of the LibreTexts platform.

- [2.5: Correlation and Causation, Scatter Plots](#) by Maxie Inigo, Jennifer Jameson, Kathryn Kozak, Maya Lanzetta, & Kim Sonier is licensed [CC BY-SA 4.0](#). Original source: <https://www.coconino.edu/open-source-textbooks#college-mathematics-for-everyday-life-by-inigo-jameson-kozak-lanzetta-and-sonier>.