CORRELATION MINI LECTURE

Correlation is a relationship between two sets of numbers (variables). It attempts to summarize that relationship, usually with a combination of a plot of data points and a trendline that summarizes the data. There are statistical means to calculate correlation, but this activity will only use this summary of the relationship. The trendline that summarizes the relationship will provide whether the relationship exists, and if so, if it is positive or negative. We will also produce a measure of fit to better understand the relationship of the variables.

# SUMMARY OF WHAT HAS BEEN ACCOMPLISHED

Since this is the third step in this activity, we should review briefly how we selected data and how we summarized the data.



1. ***Data Selection*** was accomplished by finding a variable of interest in Gapminder and looking for patterns by in the data and evaluating the type of data to ensure we could use it in our next steps.



1. ***Data Summary*** was accomplished visually with a box and whisker plot and through calculations of a 5-point summary (median, minimum, maximum, 1st quartile, and 3rd quartile).

What we have accomplished so far is to get to know the variable we are interested in. We have a better idea if there are outliers that may be interesting or need to be excluded, and we know how much variation there is in the data. Now we are ready to look for a possible cause of the variation we see in the data.


# INDEPENDENT VARIABLES

1. ***Asking a Question:*** One of the best questions for doing a correlation is to ask the question “Why?” With this question, we can ask: “Why is there so much variation?” Or “Why does that part of the world have higher values than the rest of the world?”
2. ***Choosing an Independent Variable:*** With a good “Why?” question, we can think of possible answers and use them to think of a variable that might be causing our data to be what it is. When we choose this variable, we are saying that we think it might be an **independent variable** that influences our **dependent variable** – the original variable we were interested in.
3. ***Predicting a Relationship:*** With an idea of how our independent variable influences the dependent variable, we can make a prediction. If we think more education would lead to greater acceptance of vaccines, then we could test that by looking for a positive relationship between education and vaccines. If we thought that increased education led to a lower acceptance of vaccines, then we would be looking to see if there was a negative relationship between these variables.

# GRAPHING A CORRELATION

When we graph a correlation, we are setting the two variables (independent and dependent) against each other in a scatterplot. This shows us how the two variables relate, and if changes in the independent variable are associated with changes in the dependent variable (Does more education mean a country has a higher acceptance rate of vaccines?).





1. ***Scatterplot:*** Scatterplots make coordinate points based on each individual’s (country’s)value for the independent variable on the x-axis and dependent variable on the y-axis. By plotting them all together, we can look for patterns. Labels on both axes and on the overall graph make interpretation of the graph much easier. Units of measure should be included in the axes titles.
2. ***Trendline:*** Trendlines summarize the data by minimizing the total distance from the line to each of the points.
	1. ***Slope:*** The equation for the trendline includes the slope as a multiplier of x, telling us that as x increases one unit, y increases by the slope. In this image, for every extra mmol/L of cholesterol, breast cancer cases per 100,000 increase by 41.
	2. ***R-squared:*** The trendline minimizes the total distance from the line to each point. The better it does this, the higher the r-squared. A lower r-squared means the points are further from the line and a higher r-squared means they are closer. There is no cutoff for what is a good r-squared, it is just a statistic that says this amount of variation in the data is explained by the trendline. In this case, the trendline explains 34% of the variation in the data – the overall tendency to increase. CORRELATION VS CAUSATION

Correlation does not equal causation. Just because two variables are correlated does not mean one is causing the other. For example, there may be another variable that is causing both of those variables to change.

1. ***Scatterplots and Trendlines:*** What is created when we do a scatterplot is a visual representation of correlation between the variables. We might see that as the independent variable increases, so does the dependent variable.
2. ***This may not be a Causal Relationship:*** It is tempting to say the independent variable is causing the dependent variable to change or be different. Since we do not have a controlled experiment, we are unable to say there is causation, that the independent variable is causing the dependent variable to change. All we can say is that they are correlated. For example, our scatterplot above shows a strong positive relationship between cholesterol levels and breast cancer. There is good biological reason to think that increased cholesterol levels could lead to (cause) increased levels of breast cancer. However, all we know is that countries with high blood cholesterol levels also tend to have higher rates of breast cancer.
3. ***Many Unseen Variables:*** Both of these variables could be caused by some unseen third variable, like health education. In order to look for causation, we would need to run a more controlled experiment, where we could remove or account for other variables. One of your tasks will be to describe a controlled experiment where you could test for a causal relationship.
	1. **Simple Example:** You make a scatterplot of ice cream sales against temperature. You find out that increased temperature and ice cream sales are correlation. What might actually be happening is increased temperatures in your study are actually mostly due to increased sunshine, and people prefer to eat ice cream when it is sunny. Therefore, the increased temperature is not causing ice cream sales to go up, but both temperature and ice cream sales go up with increased sunshine.

This material is based upon work supported by the National Science Foundation under Grant No.1919613. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.