Correlation Doesn't Equal Causation: Crash Course Statistics #8

Crash Course: Statistics
https://youtube.com/watch?v=GtV-VYdNt_g
https://nerdfighteria.info/v/GtV-VYdNt_g

Hi, I'm Adriene Hill, and welcome back to Crash Course Statistics.

Today, we're talking about relationships. No, not why you and your bestie are platonic soul mates, or why your cat just doesn't seem to like you, we're talking about data relationships, like how you can use one variable to predict another. Like if you can predict whether people who write in all capital letters are more likely to default on loans. Whether people drive faster after they watch *Fast & Furious* movies. Or whether blink more often when they're lying.

[Opening music]

We'll start with the simplest data relationship, one between two continuous variables, also called bivariate data. But first, we're going to need to visualize our data using a scatter plot. The scatter plot has been called "the most versatile, polymorphic, and generally useful invention in the history of statistical graphics." Impressive,and as such they are pretty much everywhere, including on your favorite news site. News outlets now have data journalists on staff to visualize and make sense of data.

To make a scatter plot of Old Faithful eruption duration and latency - which is the time between eruptions - we put one variable on the x-axis and the other on the y-axis. Then each data point is placed so that it's in line with both it's eruption duration, and it's latency. Now we can see a relationship. There are clusters, two blob-y looking groups of points, which supports our guess that there are likely two kinds of eruptions, one with a longer build up and longer duration, and one with a shorter build up and shorter duration.

Just like the histogram and dot plot, a scatter plot allows us to see the shape and spread of data - but now in two dimensions! This data is clustered, but scatter plots are useful for identifying all kinds of relationships, both linear and nonlinear.

For now, let's focus on linear relationships with a classic example - the relationship between the the heights of fathers and sons. It makes sense that a tall father would produce a tall son, but we can do better than just a hand wave-y statement.

In 1903, the statistician Karl Pearson published an influential paper - in his own journal, *Biometrika*. One section of the paper describes the relationship between the heights of dads and their male children. In this paper, Pearson fit a line through the data to describe the relationship, rather than just relying on his eyes to see a pattern.The line - called a regression line - is the line as close as possible to all the points at the same time. And note here Pearson used feet and inches in his paper so we will too.

Lines are a great way to describe a relationship because they have a nice formula: $y = mx + b$, just like you learned in algebra. The m (or slope) tells you a lot about your data. It tells you that an increase in 1 inch of a father's height, leads to an increase of m in the son's height - about half an inch in Pearson's paper. So on average dads who are 6'1 tall have sons that are about half an inch taller than the sons of fathers who are 6 feet tall. That allowed Pearson to make a prediction about the height of the son from the height of the father.

And this is why these lines are so useful - they allow us to pretty accurately predict one variable based on the value of another. The relationship between car weight and gas efficiency allows us to be pretty sure a SMART car gets better mileage than a Hummer.

One note of caution: the slope relies heavily on the units of x and y since it's a measure of how many units y increases with each increase of 1 unit in x. If I decided to measure the son's height in meters, the m (or slope) will change, even though the relationship didn't.

When we see a non-zero slope - also called a regression coefficient - it's a sign that there's some kind of relationship between our two variables, but that's pretty much all it tells us. We don't know how strong that relationship is.

For more information,we need to look at correlation. Correlation measures the way two variables move together, both the direction and closeness of their movement. You may have read articles claim that there's a positive correlation between exercise and heart health. That just means if you exercise more, your heart tends to be healthier. A positive correlation looks something like this on a scatter plot. While a negative one, like the correlation between number of cigarettes smoked each day and lung health, might look like this.

Higher values of cigarettes smoked tend to have lower values for lung health. We now know what correlations look like in general, but to understand them more deeply,we're going to take a closer look. If two variables have a positive correlation, they move in the same direction. We can see this in our scatter plot if we draw two lines across the graph - one at the mean of each of our variables - to divide the plot into four quadrants.

When two values are positively correlated, how many miles you run and the number of calories you burn, most of the points will be in the upper right and lower left quadrants. In these quadrants, the values for miles and calories burned are either both large, or both small. The more miles you run, the more calories you burn.

The opposite happens when the correlation is negative, like the relationship between vaccination rates and the rates of preventable illnesses. Instead of moving together, the variables move in the opposite direction. So, the points are mostly in the upper left and lower right quadrants where either vaccination rate is small and rate of illness is large, or vice versa. Since vaccination rate and rate of preventable illness have a negative correlation, as vaccination rates increase, rates of preventable illness decrease.

The more closely two variables move together the stronger the relationship will be, positive or negative. If the points are in all of the quadrants pretty evenly. You just have a blob or a cloud. You don't have a strong relationship.

As I mentioned before, the units of your variables can affect the regression coefficient, and can also affect the calculation of our correlation. To get around that, we use the standard deviation to scale our correlation so that it is always between -1 and 1. This is our correlation coefficient, r. Interpreting r involves two things: the sign of the number, that is whether it's positive or negative and how big the number is.

The sign will tell you whether your two variables move together (positive r), or in opposite directions (negative r). A correlation of 1 or -1 would be a perfectly straight line, meaning you can exactly predict one value from the other. Say we looked at correlation of the number of hours you're asleep vs. awake. If I know one of those values I can tell you exactly what the other one is. We all have only 24 hours a day. Even Beyonce.

As you get closer and closer to a correlation of 0, the points are more and more spread out around our regression line, and eventually at 0, there's no linear relationship at all, it's just dots.

When you look at a scatter plot, remember that you can't deduce a correlation just by the steepness of the regression line. In our earlier father/son heights example, we changed the units to meters and our line didn't look as steep, even though it's the same data. Data with steep lines can have low or high correlations. We also use the squared correlation coefficient r^2. r^2 is always between 0 and 1, and

tells us, in decimal form, how much of the variance in one variable is predicted by the other. In other words, it tells us how well we can predict one variable if we know the other.

While they won't usually give r^2 an explicit mention, you'll see articles claim things like " the ounces of soda a person drinks is highly predictive of weight", which means there's a large r^2.

You can think of r^2 as a measure of how accurate your guesses would be if you used your linear equation to predict one variable from another. If you have an r^2 of 0.7 for the cigarettes and lung health data, that would mean cigarette usage predicts 70% of the variation in how healthy our lungs are. You could pretty accurately predict someone's lung health if you knew how many cigarettes they smoked. An r^2 of 1 means you can perfectly predict one variable from the other since 100% of the variation is in one variable.

This can seem pretty obvious when you think about conversion. Like temperature in Fahrenheit can be predicted by temperature in Celsius. In this case we're not actually measuring the temperature in Fahrenheit, but it is predicted by Celcius.

So in general, the higher the r^2, the better the fit.

[World New music]

Breaking news from city hall today! The mayor has announced a plan to cut down on the number of people who drown every year. Sources close to the mayor tell us that he's seen some very interesting correlations between drownings and air conditioning usage, and drownings and Nicolas Cage movies. Or as I like to call it - air cons and Con Airs.

Both are highly correlated with drownings. Here's evidence. If we look at AC sales data over the past 10 years. And even more proof if we look at Nicolas Cage movies over the same time period. The Nic Cage data was provided to the city by Tyler Vigen. So as of today, our mayor has enacted the Cool-Cage act which will prohibit sale of air conditioners and create a Nicolas Cage task force who will do everything to prevent Nicolas Cage from starring in any movies. The Mayor assures us that because of the strong correlations she saw, as well as the strong will of our city, we will surely have next to no drownings this coming year.

The Cool-Cage act may seem silly, but we're constantly flooded with messages that equate correlation with causation. And as you've heard before: correlation doesn't equal causation. Just because two variables are related doesn't mean that one variable causes the other.

The examples the mayor uses are perfect examples of things that can go wrong when interpreting correlations. When one thing (A) is correlated with another (B), there's a few possible reasons: A causes B, B causes A, there's a third variable C that causes both A and B, even though A and B aren't related, or there's no relationship at all. it's just a coincidence.

The correlation the Mayor saw between air conditioning and drownings is probably caused by a third, unmentioned variable: heat! When it's hot people buy more air conditioners and go for a swim leading to a correlation even though there's no direct link between the two. And as for Nicholas Cage, he probably shouldn't feel too guilty about causing world-wide drownings. Sometimes two completely unrelated things are correlated just by random chance, with no causal link at all.

These correlations get called spurious correlations, and they can be hard to catch. But when the correlation is between two VERY specific things, like Nicolas Cage movies and all drownings in 3 feet of water when a dog was present, you should be suspicious that someone tried every weird subset of data until they found a relationship.

Before we finish with correlation, I just want to warn you: r and r^2 aren't everything: it's important to look at a scatter plot of data when you can. These are the "Datasarus Dozen" these very different plots all have the same correlation, but we can see that the relationships are completely different.

Correlation is an important piece of the puzzle when you're looking for a linear relationship between two variables. It goes above and beyond the $y= mx + b$ and gives us information about how well that line explains the data. Understanding the relationships between variables and events helps us predict what things are going to happen in the future, and also reflect on why things occurred in the past.

A correlation could help you predict how much money you'll make after years of working your way up as a lemonade salesperson. Or if watching that next *Fast and Furious* movie in the theater might encourage people to speed. According to an analysis by a Harvard Medical School professor Anupam Jena, those two things could be related. Relationships are important the human-kind and the data-kind. Correlation allows us to better understand relationships between data. And maybe also the data of our relationships.

Maybe you can find correlations between the amount of time you spend at work or school and how much affection your cat shows you. Mr. Fluffy misses you. Thanks you watching. I'll see you next time.

Crash Course Statistics is filmed in the Chad and Stacey Emigholz Studio in Indianapolis, Indiana, and it's made with the help of all of these nice people. Our graphics team is Thought Cafe.

If you'd like to keep Crash Course free, for everyone, forever, you can support the series at Patreon, a crowdfunding platform that allows you to support the content you love. Thank you to all our patrons for your continued support.

Crash Course is a production of Complexly. If you like content designed to get you thinking, check out some of our other channels at complexly.com.

Thanks for watching.