Hi. This tutorial covers correlation. Let's just start with a situation to motivate this idea. So I'm interested in knowing how driving commute time is related to the distance a worker lives from the workplace. I asked nine colleagues to provide me with the distance they lived from work rounded to the nearest mile. I also asked them to keep track of their driving commute times in minutes over two weeks and provide me with their average commute time for those two weeks, and that's going to be rounded to the nearest minute.

OK, so here is the data that I was able to get. These were distance from the workplace, and then this was the commute time. These were in miles. These were in minutes. So one person lived 6 miles from work. It took them 7 minutes to get here on average. This person, they lived 12 miles from work. It took them 10 minutes. And you can see that the nine data values here.

So here is the data as a scatter plot. OK, so the x-axis is the explanatory variable, which is the distance from the workplace. So the distance explains the commute times. So commute time is the response variable. And you see the nine values there.

OK, so let's consider the direction, the strength, and the form of this association. OK, so direction I would say is positive. As the distance increases, the commute time increases. Strength, I would say that's probably moderate. We can see that they seem to be pretty clustered about a line, but I would say it's not a super strong relationship. And then the form I would say is linear. So I would say that this is a strong positive linear association.

Now, if the form of an association is linear, there is a quantitative way to measure the direction and the strength of an association. It is called the correlation coefficient. So let's make sure we first of all define the word correlation. So correlation is the tendency for two variables to increase at the same time, which is called a positive correlation, or whereas one value one variable increases, the other variable decreases. That's called a negative correlation. Correlation, we're talking about only linear associations. Now, a correlation coefficient is the quantitative way to measure the strength and direction of a linear association. Correlation coefficient is often abbreviated as r, so lowercase r.

So let's take a look at the formula for the correlation coefficient. So we can see that r equals, now, it's the sum of z sub x. So really, what that stands for is the z-score for each x value times the z-score for the y value, so z sub y. It's the sum of the product of those two, so we're going to have to find the z-score for each x and y, multiply them together, and then add them all up. Once we have that sum, we divide by n minus 1.

So recall that the z-score formula is x minus x bar over the standard deviation of x. And then z sub y is going to be the same, but these are all going to be y's. OK, so basically, what I'm going to do in this table here is I'm going to put all of the z-scores for x's here and all the z-scores for y here. And then, I'll multiply those two columns together, add up the product of the two columns, and then divide by n minus 1.

I'm going to do most of this in the calculator, but I'll show you how to do it by hand also. So I have the data in L1 and L2. So in order to get the z-scores for x and y, what I need is the mean and standard deviation of all of the x values and the y values. So what I'm going to do on my calculator is do what's called two variable statistics, and I'm going to do it on my two lists of data here.

And I'm going to write down the important information that I'm going to need. So x bar is 11. So that means, on average, my colleagues lived about 11 miles from work. s sub x is the sample standard deviation of the x's. So that's going to be about 3.968.

Now I need my mean and standard deviation of the y. So y bar is going to be 12. That means, on average, it takes about 12 minutes for my colleagues to get to work. And then s sub y is 3.873. So these calculations can also be done by hand using the mean and standard deviation formulas, but I'm going to use these values.

So now, what I'm going to do is go back to my data. And I want to find the z-score for the first x value. So my first x value is 6. So I'm going to calculate x minus x bar over s sub x. So 6 minus 11 divided by my standard deviation, which is 3.968. So that ends up being about negative 1.260.

Now, if I wanted the z-score for my y value, I would need to do 7 minus the mean of y, which is 12, and divide by 3.873. And that's going to give me about negative 1.291. So I need to do that for all of my x and y values. So I'm actually going to do those in the calculator lists.

So I'm going to go up and put a formula into my third list here, which is going to be my x values minus the mean of the x's divided by the standard deviation. OK, notice my negative 1.26 matches what I got by hand. Then, in L4, I'm going to do the same thing, but for the y's now. So L2 minus 12 divided by 3.873. And notice again, my y z-score matches the one I had had there.

Now, my next step is to find the product of these z-scores. OK, so if I just wanted that this product-- so I'll make a new column over here, which is the product of zx and zy. So if I did that first one by hand, I'd get negative 1.260 times negative 1.291, and that gives me about 1.627. So to get all of these numbers using the calculator, I'm going to go into my fifth list, and I'm going to take my third list times my fourth list. So that's the product of my z-scores. Hit Enter there, and we can see that that number is pretty close to the one that we calculated by hand.

So these give me all of the products now. So what I'm going to do is find the sum of that new list, which is going to be zx times zy. And I'm going to do that by using the sum function. So I'm going to do the sum of that fifth list that I had. So this is going to be about 5.986.

So now, to finish up my r calculation, r is going to equal 5.986 divided by n minus 1. So n was equal to 9, so I'm going to end up dividing that number by 8. So I'm going to go divided by 8, and that gives me a correlation coefficient of about 0.748. So r equals 0.748.

And then a couple other things about r now that we have some experience calculating it. r will always be a number between negative 1 and positive 1. The closer r is to negative 1 or 1, the stronger the linear relationship. And r has no units.

So just to make sure we have a good understanding of what the extremes look like, if we had this association, r here would equal positive 1. That's a perfect positive correlation. If we looked at something like this, points right in a row going down, r would equal negative 1. And if we just had a cloud of points with no association, here, r is going to be approximately 0.

OK, so these are just three examples, kind of the extremes of your correlation coefficient. All right, so that has been the tutorial on correlation. Thanks for watching.