Robert_S3_L06

Thu, 2/17 4:01PM 11:02

SUMMARY KEYWORDS

outliers, scatterplot, relationship, quantitative variables, variable, correlation coefficient, correlation, linear, observations, number, axis, categorical variables, conditions, line, linear equation, mathematical definition, increasing, coefficient, straight line, negative

SPEAKERS

Robert McKeown

R Robert McKeown 00:05

Hello, and welcome. Now that you know what correlation is and how to calculate it, I want to emphasize to you the importance of understanding the three conditions, when the correlation coefficient gives an accurate measure of the strength of the relationship between two variables or two series of values. Correlation is a linear measure. So it's going to work best when the relationship between perhaps the X variable Y variable, if you want to think of it like that, or really, any two quantitative variables are linear. I'm going to show you what that looks like on a scatterplot. And I'll show you what that is not on a scatterplot. We'll look at some examples of nonlinear relationships. For the correlation coefficient, to be accurate, there are three conditions that must be satisfied. The first is that the variables themselves are not categorical variables. That is that they are quantitative variables. What's a categorical variable? Well, those were the kind that might be say yes or no answers to questions. It could be something like a Likert scale. Although it's a number, you only have options from one to five. Where one might be, you know, very unhelpful. And five might be very helpful. That sort of thing. So although they're although they're coded and numbers, they represent statements from people and their frame of mind. What's another categorical veil, variable, family status? Married, single, other, etc. So you only want to use the correlation coefficient that we've looked at in earlier videos, you only want to use that correlation coefficient with quantitative variables. There are other types of measures we can use for categorical variables, but we're not going to discuss them in this video. We want to make sure the relationship between the quantitative variables is indeed a linear or straight. That comes from our really our definition of what the correlation coefficient is how it's calculated. And we'll look at number two in more detail in a moment. Finally, you want to avoid outliers. Outliers are extreme, you know, extreme. I don't want to say extreme values. But I will say extremely far from the mean. And they may have other unusual characteristics in terms of their relationship between the X and Y variables, or the two variables or two series that we're interested in understanding. The tricky thing about outliers is there is no one specific cleared, mechanical mathematical definition for an outlier. So as a practitioner using statistics statistics, it's always identifying outliers is more of an art than it is a science you have to use judgment. And then if you do identify an outlier, what to do about the outlier also requires judgment. That is beyond the scope of the video that we're looking at today. Today, I just want you to be aware that outliers can dramatic sometimes dramatically change the

correlation coefficient to either make it much larger or much smaller. And so it can make a strong relationship look weak, and it can make a weak relationship look strong. Now, number one, you can check this by looking at the data specifically the values in a series the answers, if the data is yes and no, then it's clearly not going to be a quantitative variable. For number two, and three, we can check these using a scatterplot diagram. In fact, one of the best ways to make sure that the relationship is straight and that they're no outliers is in fact with a scatterplot diagram. And we're going to do that right now with some prepared examples. We have two scatter plots presented in front of you. On the screen, you can see scatter plot a and scatterplot B. Now, the first thing we might ask ourselves is, are these quantitative variables. And if we look at the Y and the X axis, it looks like and specifically scatterplot. A, it certainly looks like they're quantitative variables. And they're kind of randomly distributed within the range. The next thing we might ask ourselves is the relationship linear. And if you look at the distribution of our observations, we can see that as we're moving from left to right, on the X axis, the observations themselves are increasing on the Y axis. And so this looks like it could be linear. Now that's that's sort of the direction of the relationship. But does that mean it's linear? Well, not necessarily. The best way or my preferred way to show that this is linear is can I stick a straight line right down the middle, and have it be relatively close to all these observations? Let me show that to you. Again, if I can draw a straight line, somewhere on this diagram, such that all the observations are close to that line. Like so I'm doing with a thicker line, then the relationship seems to be linear, it seems to be moving in a straight line, as we increase our X value, the Y axis, or the Y value is increasing proportionately. And the proportion by which y changes is the slope of our linear equation. Remember, Y is equal to mx plus b, which is one of the forms for writing out a linear equation. Lastly, we want to make sure that there are no outliers. So an outlier would be an observation far away from this line, often sitting out there alone, or with maybe one or two. Close by observations, an example of an outlier might be an observation out here, or maybe something like where we have our x axis variable is equal to zero, and our y axis variable is equal to eight. So outliers would be something like that off into the upper left quadrant. But since we don't see anything like that, deleted, doesn't seem like we have any outliers to worry about here. Turning our attention to scatterplot B, we can see that as the X axis variable increases, our Y axis variable is decreasing. And so this is a negative relationship or correlation. While scatterplot, a was a positive relationship, ship scatterplot. B is a negative relationship, and we'd expect the correlation coefficient to also be negative. Similarly, to scatterplot a, although the relationship is negative, I can draw one of these lines through the observations and have them all be relatively close to the line. And so the relationship to me looks like it is indeed, linear. The observations are just as close to the line up here as they are down here. That's kind of what I mean when I say the observations are close to the line. In this case, I mean that it's there about the same distance from the line throughout the entire scatterplot diagram. Scatterplot A and scatterplot B. Both these scatterplots satisfy those three conditions or three situations are factors that we're looking for, in order to believe that the correlation coefficient is doing a good job of measuring the strength of the relationship between two variables. In the next video, we'll turn our attention to situations where those conditions are violated. And this can be a really powerful way to learn. It's neat to see what what these scatterplots look like when correlations working but also helpful to understand what do they look like when the correlation coefficient is not going to be an accurate representation of the relationship between two variables.