

module_prob4_lecture4

Fri, 2/18 12:51PM 15:37

SUMMARY KEYWORDS

correlation, variance, covariance, tilde, calculate, negative, relationship, height, correlated, correlation coefficient, high, measure, squared, uncorrelated, birth rate, standard deviation, weight, multiply, clip, point

SPEAKERS

Sumon Majumdar



Sumon Majumdar 00:05

Since the last clip, I introduced you to the idea of covariance of measure which can be used to look at the relationship between two variables x and y . So we've covariance is great and telling you whether the relationship is positive, or whether on average the relationship is negative, it doesn't quite tell you how strong that negative or positive relationship is. So in our example, we got a covariance of negative 2.8. So that means on average, the relationship between X and Y is negative that means when x is high, y is more likely to be low, and vice versa. But how strong is this relationship is when x is high, is it very, very likely that y is going to be low? Or is it only moderately likely that y is likely to be low. So together that I'm going to introduce a related measure, which is called correlation. And this is a measure which is very widely used also at statistics. So do try and under understand this concept, so what is correlation. So the correlation between x tilde and y tilde, and this is denoted by the, the Greek letter rho. So what it does is that it takes the covariance the thing that we've already learned how to calculate and divide it by the square root of the variance of X and the variance of y . So it takes the, again, the covariance, which is σ_{xy} , and divide it by the square root of the variance of x . So the square root of the variance of X is its standard deviation, and the square root of the variance of y is its standard deviation. So you can also write this equivalently Oh variance divided by the standard deviation of x and the standard deviation of y . So why, why is this? So this is called the correlation or also called the correlation coefficient. And why is this important and useful? It's because this correlation coefficient only lies between one and negative one. So the biggest value that it can take is one, the smallest value that it can take is negative one. So why is that good? So that's good, because if you get a value, which is close to one, right close to the maximum, that means that's a very strong positive relationship between one or the other hand, if you get a correlation coefficient of minus point nine, that means it's a very strong negative correlation between x and y . So what that means is that when x is high, it's very, very likely that y is going to be low and when y is high, it's very, very likely x is going to be low. So that's if you get a correlation coefficient of negative points. So, the correlation coefficient, the closer it is to one or negative one, that shows that stronger is the relationship whether positive or negative between x and y . So, on the other hand, if x if the correlation coefficient is zero, that that means x and y are uncorrelated. So there is no relationship between x tilde. So what's an example right. So, suppose we are looking at let's say, the price of houses and the price of fish.

So, this is x tilde this is y tilde. There these two are probably uncorrelated, right when the price of fish is higher low has no nothing to do with the price of houses. So that would be an example where have two variables which are uncorrelated. On the other hand, if the if the correlation coefficient is positive. That means x and y are positively correlated. So for example, what's the relationship between let's say x tilde as the price of houses. And y tilde is rent. So these two are likely to be positively correlated when housing prices are high, the rents are also high. Similarly, if you look at height and weight, empirically, some researchers have found a correlation of about point eight between height and weight. So that's a strong positive correlation between height. In the extreme, if x and y if the correlation between X and Y is exactly equal to one, that means x and y are perfectly linearly positively correlated. So what does that mean? So, what that means is that if you in that case, if you plot x and y , there will be in an exactly straight this will be the case where ρ is equal to one in a straight line and upward sloping straight. In fact, in this case, you will know that x tilde will just be e plus b times y tilde, where b is something. So for example, if you look at height in meters and height in feet, these two are perfectly correlated, because one is just a multiplication of the other, or your weight in kg's weight in pounds, these two are perfectly positively correlated. On the other hand, right, if if you have ρ equal to negative one, this means these are again, in a straight X and Y are in a straight line, but they are in a downward sloping straight line. So this is the case where X is equal to E minus B . So more generally, if the correlation coefficient is negative, that means x and y are negatively correlated. So what's an example of that? So for example, when price of any product is high, demand for it is usually low. So price and demand are usually negatively correlated. Another example is, if you look at growth rate of an economy and birth rate, these are usually negatively correlated, if an economy is growing very fast, birth rates are usually so So again, the important thing is that the correlation coefficient is between minus one and positive negative one and positive one, the closer it is to minus one, something like point seven, five and lower, that means it's a very strong negative relationship. On the other hand, if it's positive seven, five or higher, that means it's a very strong positive relationship. Now, one other feature, which is very interesting useful is that if you take x and y and you multiply x by something like a by a constant C , or you multiply y by a constant D , or if we add constants a and b , that does not change the correlation. So what is that useful? So for example, if you remember earlier, we talked about a relationship between height and weight. Now you can measure height in lots of different ways. You can measure it by kg's, and pounds, and so on. You can measure weight by meters, feet, yards, anything. What this tells you is that whichever unit you choose, it's not going to change the correlation. Because pounds is just a multiplication of height by a constant, the weight I'm sorry, these two that mean kg in pounds, that sort of weight. And this is for height. So the point is that if you multiply if you measure height in feet or meters, or weight in kg or pounds, that does not change the correlation. So the correlation is unit free. So, let me show you how to calculate a correlation. So this is in the same example that we had done before, where remember we had two possible values of x and three possible values of y . And these were the probabilities, the joint probabilities. And we're already calculated the covariance to be negative 2.8. Right. So how do we calculate the correlation? So, we already have the covariance, so the covariance is negative 2.8. Now, what we need to calculate is the variance of X and the variance. So, how do we calculate the variance of x ? So, remember, when we had the formula from the previous module, that the variance of X is expectation of X squared minus expectation of x squared. So, we have calculated before that the expectation of x that's 14, the only thing we have to calculate now is the expectation. That's correct. So, how do we do it? So, see, let's look at X squared, X takes the values 10 and 20. So, if I take 10 squared, multiplied by the probability, that's point six. And the other possibility is x takes the value of 22. This is 20 squared, multiplied by its probability point four. So this gives you an expectation of x squared. So this is going to be going to be 220. And now if you plug it back in here, you get saved for plug in back here. 20 minus 14 squared, that gives

the variance of x , which is going to be so replace here 24. So that's the that's a variance of x . Similarly, we need to calculate the variance of y . Again, if I do variance of y , that's going to be expectation of y squared minus expectation of Y , all things squared. This we have already calculated before, which was 2.2. Now to calculate the expectation of y squared, we need to take the possible values of y . So one squared, times point two, two squared times point four and three squared times point four. If you do that, that'll be 5.4. And if you if you do this calculation, this variance is point five, six. So I'm going to then replace this by point six. And once I do this calculation, that gives the correlation between x and y , this will be a negative. So what this tells me is that between the x and y , there's a fairly strong negative relationship. So when x is high, y is more likely to be low, and when y is high, x is more likely to be. So this is how you would calculate. So let me end this clip with a clicker question. So he stopped the video and he attempted. So hope you had the chance to do the clicker question. So what it asks us which one of the following correlation coefficient measures the strongest relationship between random variables x and y , so there is negative point eight negative point four, point three, point six. So so this is the One that has the strongest negative relationship. So, point six is a positive relationship. But if you look at which is the strongest, that's point eight, so the correct answer to that is so let's start the clip here. So so far I've talked about covariance and correlation. And in the next clip, I'll be talking about the main topic of this whole module, which is continuous random variables.