Robert_S3_L13

Thu, 2/17 4:01PM 19:24

SUMMARY KEYWORDS

sample covariance, covariance, column, income, calculate, education, cell, deviations, excel, observations, parentheses, highlight, mistake, correlation, select, number, error, calculation, decimal places, formula

SPEAKERS

Robert McKeown

R Robert McKeown 00:07

Let's take a look at the questions on the screen in front of us. We have three columns. The first is the question number. And then we have a sheet where we can find the data that we are meant to manipulate. And finally, we have a column that has the actual question in it. And so the first question we want to answer is this one right here. And the question is asking us to calculate the covariance and correlation and the correlation of determination between education and income on the sheet, named education. So let's click on the sheet named education. Looking at the sheet, we have a title years of education and income and Canadian dollars, here we have years of education. So 12 years would be that would be equivalent to a high school degree. And if we have 13 years of education, that would be a high school degree plus one year, post secondary education. And then we have associated income listed here. And so we have information on each individual's income, and their highest level of educational attainment are defined as how many years of education they have. Now, our first question wants us to calculate the covariance, and the correlation and the correlation determination. So we want to start with the covariance, we need the covariance in order to calculate the correlation. And we need the correlation to calculation how to calculate the correlation of determination. So I'm going to start by creating two columns, and we'll call them education. deviations from the mean. And we'll create a second column, notice that I double clicked on this little boundary between the C column and the D column. And that made the column the width, such a will fit the largest series of characters of any of the cells in that column. Next, we're going to create income, and we'll call it deviations from the mean, I'll just write deviations to save a little bit of space. And now we want to calculate these deviations from the mean. So I'm going to highlight the cell C 3, a hit the equal sign, and I'm going to select a three, which is the first cell with information about educational attainment. And I'm going to subtract the average because in Excel language, the mean is called the average. And I'm going to take the average of all the information in cells A 3, all the way down to the end of the series of information, which is in cell A 102. I'll close my parentheses. So you can see that we've got the formula up here. Now I'm going to make one alteration to it, I'm going to pin or lock the column and row coordinates, so that the mean won't change when I use this formula to populate other cells. And to populate the other cells. Of course, I'm going to click in the bottom right corner here. And then I'm going to left click, and I'm going to hold it and drag it all the way down and

populate all these other rows. Then I'll hit control Up, Up Arrow, Control up arrow to take me back to the top. Now if we wanted to, we could also freeze the top row. So that wherever we go down if we go further down the worksheet, we'll always be able to see the names of columns. So why don't I do that? Next, we want to find the income deviations from the mean. So we're going to hit the equal sign in cell D 3 equals two B 3 minus Y minus the average of all income. This time, I'll click the cell B 3, then I'll press control shift and the down arrow, and that will highlight. All right, all it'll take. It'll highlight all the information up until there's a cell with no information in it. Close the parentheses you can see the formula where you same up here, I press enter. And now we've got our deviations from the mean in income. So at income of 87,319, has a deviation from the mean of 14,284. Now I'll select a highlight the cell, click the bottom right corner and drag down. Notice this time the column names remain visible. And now we've got all this information, maybe I want to change the format, so it doesn't look quite so I don't know quite so messy with all those decimal places. And we'll turn it into a dollar sign, and a dollar value while we're at it. So we're gonna have our cents and our dollars. Now, going back up to the top, we now have our deviations from the mean, I guess going back down to the bottom, I hit Ctrl, arrow down, we can just verify that we calculated this correctly. I'm going to type in D 3, to D 102. There's a notice that there's a colon separating the two. And it looks like I made a mistake. And what mistake did I make I forgot to add dollar signs. So let's go ahead and highlight all the way to the top, I might have to just highlight it like so. And we'll clear I'm going to right click on the highlighted cells. And I'm going to choose Clear Contents. And then I'm going to go back into the cell D 3. And I'm going to add those dollar signs. So notice that it's not a bad idea to make sure that your calculations are correct, wasn't obvious to me that there was a mistake. Now we're going to sum up all these deviations from the mean, from cell D 3 to cell D 102. I press enter and get the answer that we're looking for, which is which is zero. Now let's do the same thing with education and its deviations from the mean. I'm taking the sum of all the values from C 2 to C 102. I press Enter, and we get this value here. Now, you might say Oh, well, maybe we made a mistake in the formula. But looking at this, this is scientific notation, specific to computer programs. I often don't I don't usually don't see it anywhere except in computer programs. And it's actually telling us that this is a really small number. So this is really equal to maybe if I format this cell here. And I choose a number with zero decimal places, you can see it's a really small number, it's negative 1.42109, multiplied by 10 to the power of negative 13. 10 to the power of negative 13 is a very small number. This is what I was taught to call machine zero. I'll type it out machine zero, sometimes computers and having a very small rounding error. And this is the result or possibly ran an error other kind of calculation error. But essentially, we're really, really close to zero here. So that seems like the calculation for deviations from the mean. And education is fine, I'm not worried that we've made a mistake there. Now that we have our deviations from the mean, we can go ahead and calculate the covariance. So I'm returning to the top of the spreadsheet. And part of our covariance calculation is to multiply the deviations from the mean and education by the deviations in the mean and income. So why don't we do that I'll type in an equal sign in cell a three, and then I'll select cell C three, this little star button which is Ctrl, eight on my keyboard and multiplied by d three. Now I'm going to highlight the cell that we've created, and I'm going to drag, drag it down and populate all the cells up until the end of the data set. So we're going to perform the multiplication for each of our observations, now we've got some little number signs here, we can get rid of them by double clicking on the boundary between column E and F. And that will give it a more size so that we don't have any sort of ugly, hidden values like that. Now we're ready to calculate our covariance statistics. So I'm going to sum all the values in the E column. So from E 2 to E 102. Or I guess that's a three, excuse me, E 3 to E 102. And make the column a little bit bigger, so we can see it, we've got this rather large number there. But of course, to get the covariance, we are going to divide it, and we're going to divide it by the number of observations, n, but the instructions didn't tell us we were to calculate the COVID population

covariance or the sample covariance. So I'm going to err on the side of caution, I'm going to take a conservative estimate. And I'm going to calculate the sample covariance. If the instructions don't tell us whether to use the population or sample, I'm going to use the sample and I suggest that you do as well. Now, to calculate the sample covariance, we need to know how many observations there are. And we can count that in a number of ways. When if I go back up to the top of the spreadsheet, and I highlight the A column, I'm going to right click and choose this insert, button. And Excel knows that I want to add a column to the left of the A column, which is going to push what used to be in the A column into the B column. And I'm going to call this column observations. And I'll make it a little bit larger, so we can read the name. And we've got our first observation or second observation. If we think back to the lecture slides, that could be the first value in a series the second value in a series. But in statistics, we often call these observations. And I'm going to just highlight the one and the two and start clicking and dragging down, Excel will know that I want the interval to just be one for each observation. Gonna go down to our last observation, which is itself 102. And there's a way to find out how many observations we have very simply, hopefully, in a way that doesn't make a mistake, we don't make a simple calculation error. Now, if I want to calculate the sample covariance, and I will move this cell so that it's right adjusted right aligned, just to make it a little bit easier on the eyes. And I'm going to take this sum that we have in cell now Excel F 103. And I'm going to divide by open parentheses, the 100 minus one. Of course, maybe you want to just select the 99. But we're going to do this in a systematic way. Systems are good when you're working with Excel or computer programs, because we really want to avoid making mistakes. because mistakes are very, cost a lot of time they're very time costly. So we want to avoid errors, try and do things systematically. And it says that the covariance, or the sample covariance is actually $55,013. There is an easier way to do this, we could have just called on the excel covariance formula covariance sample, and we could select the array. So we know that education, it goes from B 3 to B 102. And income goes from C 3 to C 102. And notice I separate C 3 and C 102 with a colon. Make sure your parentheses are in the right place and press Enter. And we can see that changing the format make it similar that our covariance or sample covariance calculation that we did the long way is equal to the short way that Excel did for us. And so, both of these are sample covariance. Now we want to calculate the correlation and the correlation is equal to the covariance divided by sx as why so we need to have the sample standard deviation of education And our sample standard deviation of income. We can do this by moving to cell F 107. Maybe I'll just scroll down a little bit, so it's in the middle of your screen. And we want to select not F 108, we want to select our sample covariance, this forward, slash backslash, backslash. And we want to create an open parentheses. And we can just add in the Excel functions directly. So if I press S, T D EV sample, we want the standard deviation of education, which is B three to be 102. And we want to multiply that by the standard deviation, sample standard deviation of income, which is in cell C 3, to C 102. And make sure I've got two parentheses to close it, I have to close each function. And since I had one, open parentheses, and I need to close a standard deviation function, we end up with two, as you can see highlighted there too at the end, if I press Enter, I don't get any errors. So we did it correctly. And here's our correlation coefficient. Let's change the formatting so that it's a little bit easier on the eyes. And when we round, we find we have a correlation coefficient of 0.6. And that's the answer to question one. I suppose there is one last part to question one, which is the correlation of determination, our R squared. And our R squared is just going to be equal to the correlation squared. And so I'm going to use this circumflex that tells Excel to raise F 107 to the power of two, I hit enter. And got a really big number. Let's again, let us change the number of decimals to little easier on the eyes. And we find that our correlation determination is equal to 0.35 when we round to two decimal places. Now question two asks us to interpret the result. And so there's a couple things that we can say about the relationship between education and income. The first one is we can say that there is a I'll call it a

moderately strong correlation between education and income. 0.6 is a not the strongest correlation, but it's reasonably strong. What else can we say we can also say there. Alright, could say that 35% of the variation. And education, or I should say, an income is explained by the variation in education. So looking at this measure of R squared, we can say that there 35% of the variation income is explained by the variation in education. The rest, we don't know. We don't know. What is determining the income of individuals. But that 65% We don't know is not being explained by education.