Robert_S2_L15

Thu, 2/17 4:06PM 22:00

SUMMARY KEYWORDS

excel, calculate, sample variance, sample standard deviation, population variance, ages, standard deviation, variance, bin, cdf, cell, deviation, add, histogram, frequency, decimal places, function, called, remember, prime ministers

SPEAKERS

Robert McKeown

R Robert McKeown 00:06

Hello and welcome. In this video, you're gonna get an opportunity to calculate the variance and the standard deviation using Excel, we'll do it in two different ways. We'll build it up without using any of Excels functions. And we'll do it by looking at those deviations from the mean and then squaring them. We'll also call on Excel functions that skip all those steps and can make it very simple. To calculate standard deviations and variances. We'll look at the difference between sample standard deviation and variance and population standard deviation and variance, we're also going to calculate some probability distributions, some cumulative distribution functions, and we're going to call on a very useful function in Excel called frequency. And then finally, we'll create a histogram. This time, we'll create a histogram using the analysis tool pack inside Excel. And so I'll show you how to install that and get that working. That said, let's get started. And let's get down to it and learn something about how to perform statistics using Excel. Question number four is asking us to calculate the variance and standard deviation. I'm going to do this in two different ways. First, we'll calculate the variance and standard deviation using the formulas that we learned during the lecture. Then we'll use the functions that Excel has available to calculate the variance and standard deviation. To get started, I'll give myself just a little bit of room here, I'll select the rows 13, and 14. And I'll create two new rows. So I won't feel too cramped. I'll leave our previous question information there. We don't need that for now. In fact, I'll show you we could hide it if you were interested, we could highlight the columns E, F, G. We right click on that we can go to hide, and we just hide the information. If it's unseemly or distracting in some way, there, it's gone. We don't have to worry about it anymore. Why don't we calculate deviation from the mean. And we could get started with that before we actually start calculating the variance and standard deviation. So in order to calculate the deviation from the mean, we're going to take the observation, which is, in this case, 40 years of age for Joe Clark. And that's in cell C 16. And we're going to do minus remember that the mean, the function that calculates the mean in Excel is actually called the average. And we'll select our ages. And I'll close the parentheses and then hit enter. And here we have our deviation from the mean, all without any intermediate steps, just having the function embedded in an abroad or mathematical operation. How much precision do we want, I think two decimal places is more than enough. That's really just a formatting issue. It's just what we see when we're looking at Excel, all the decimal places are still stored within Excel. By

changing that option there, I didn't actually change how much precision is stored in Excel. So there's still many decimal places saved in that information just a little bit easier on the eyes. Now I can now that I've got a formula here that I want to keep using, I guess, I do have to make one change to it, I don't want the I don't want the average change, I want the average to stay the same, make sure it keeps trying on the column, C 16. To see 38 We don't want that to change. And that's what the dollar signs will do. It'll pin it for us. Now I can populate the rest of these cells. And here we get deviations from the mean. So remember, the mean is 55.56. And so Prime Minister topper is quite a bit older than the average. So as deviation from the mean is positive. And Joe Clark is quite a bit younger than the mean and his deviation from the mean is negative. Now we can square, so let's go ahead and square the deviations from the mean. That's what we're going to need to do to catch to like the variants, remember squaring it's really handy, because it's going to get rid of those negative signs and give us really a whole bunch of positive numbers that we can add up and divide. So how many decimal places do we want here? Oh, maybe one is enough, we don't need to see any more precision than that. Before we go too much farther, remember that the deviations from the mean, the sum of them all, is going to be equal to what? It's going to be equal to zero. And that's why we essentially, that's why we need to square them. So that we have a good grasp of how far away they are from the main, if we didn't square them, it'd be hard to take advantage of the information they contain. Now that we've got the square, I can just populate all these different elements, all these different cells. And I could sum up all the squared deviations from the mean. And that gives us 2193. And if I wanted to calculate the variance, well, maybe I'll do it up here. Say the variance is going to be equal to the sum of the squared deviations from the mean, divided by the number of observations, which is 23. And we've got a variance of 95.38. course this is our population variance, we haven't adjusted the denominator. And if we want to calculate our population, standard deviation. And I can just maybe cut these and move them up a little bit using the Ctrl X shortcut. And I'll take the square root. So we can use the square root function in Excel, like so we see that the standard deviation in the Prime Minister's age, or Prime Ministers ages is 9.76, or 9.8. Now that's the harder way or the full way of doing it, it's good to be able to perform those calculations in Excel because you're likely to need to do that for your assignments, for essays for research that you're doing and in your career, if we want to use the function in Excel, we can do that. And I can type equal and STD, and we get standard deviation. We've got both the sample and the population standard deviation. So if I want to verify that our methodology was correct, I can calculate the population standard deviation, all it needs is the ages, which I will add in and then I hit close parentheses and enter. And we get oh, I guess I put it in the wrong cell, we get 9.7661. And if I format that cell, with one decimal place, we see that the calculation we did the long way is equal to the calculation that Excel performed. How about the variance, we can calculate the variance as well. I hit the equal sign and type in var. If we want to calculate the variance, the pop- and we want to calculate the population variance. We choose var dot p. And then all we have to do is select the data the ages of the prime ministers, close the parentheses, hit enter. And there we have our population variance. And you can see that Excel is returning the same population variance that we did when we calculated it together. Question Five is asking us about statistical bias. What's the difference between the sample and population variance and standard deviations? So I think we're different Asking us is let's calculate the sample standard deviation and the sample population variance. And let's see what the difference is. When we compare that to the population. Now there's a few things we could do, we could go back to our regular standard deviation calculations, I could, maybe what we could do is this, I'll copy these cells down here. And then I'll change the name from population to sample. Sample. And we had a bit of a problem because J 39, A 38, should be J 39, A 38. There. And A 38. Remember, A 38 is this cell here, the number of observations 23. So if we want the sample, if we want the sample variance, we can change the formula, add in our parentheses, and then add an a minus one, or subtract a minus one. And we can see that when

we do that, our sample variance is 99.7. And the sample standard deviation is 10. Next, we can ask Excel to use its functions to cap to calculate the sample variance and the sample standard deviation, just to make sure that we haven't made a mistake, doing it kind of a long form using the formulas. So I will call on variants, but this time it will be var dot S. For sample variance, choose the Prime Minister's ages close the parentheses, and we get a sample variance, the same 99.7. If I take the square root of at this time, I'll do equal, I'll choose the cell K 13. And then I will do a circumflex. Point five, that's equivalent to square root. And we see that sample standard deviation is equal to 10. So if we want to know the difference, the difference is and the variance is going to be 4.3. So the if we're using the population variance, when we should have been using the sample variance, the population variance is going to be 4.3. Too low. It'll give us a under estimate of the true variance. And the standard deviation is going to be 0.2. to low. If we're using the population standard deviation, what we should be is in the sample standard deviation. So question six and questions seven, maybe I'll highlight both of them in bold right here. These questions are asking us to essentially create a histogram, but also to take our ages and put them into bins. So we want to put them into groups of ages 40, to 45, 45, to 50, 50, to 55, and so on. Also, here's a little trick, we can hide rows. So I'm going to select cells, or I should say rows 1 to 6. And I'm going to right click on them, and then I'm going to hit this hide button. And now, those questions we already answered are no longer on the screen, but they still exist. They're still there, but they're hidden. Now to do this, I'm going to show you an interesting command. Now one thing we need to do is we needed to find the bin sizes. So I'm going to define the bins as follows, I'm going to have a bin called 45, I'm just going to use the higher end of the range, gonna have a bin called 50, 55. And if I highlight these and I drag down, excel is smart enough to know that I want a width of five between each of each of these values, difference of five. And we can call this the bin- bin sizes. Now, what's interesting here is we want to calculate the frequency. So what we really want here is for each bin size, we want to know the frequency how many prime ministers were of that age in history, and it's kind of a hard thing to program but there is a function Excel called frequencies. So if I hit equal, and I start typing in frequency, see, it says calculates how often values occur within a range of values, and then returns a vertical array of numbers. So let's select that function. The first array is the data. So we're gonna select the ages, I'm going to press comma, and then it wants the bins array or really wants the bin sizes. And I'm going to select the bin sizes, they're going to close the parentheses, parentheses and hit enter. And we see that we get a number of frequencies. Now we don't have anyone over 75, up to 80. So we don't really need these bands here. If we sum up all the frequencies, we do get 23. And remember, Canada has had 23 Prime Ministers. And so now we have the frequency associated with the bin sizes. And we can go ahead and calculate the PDF and CDF. So the probability distribution function, or just the PDF, or the probability of having a Prime Minister 45 years of age or less, is going to be equal to three divided by 23. And we can change the formatting here to maybe turn it into a percentage with one decimal place, so it looks kind of nice. And we can, oops, I made a mistake. Better undo that with CRTL Z, I forgot to add in my dollar signs here. So we don't want to don't want the total number of observations changing. And I can pop, now I can populate the rest of the cells properly. And if we sum up the entire PDF, we're good to get 100%, which is exactly what we want. So using the history, as an example, if we wanted to, on, maybe we wanted to try and use history to predict the future. And we could say something like the probability of having a prime minister, over 45 years of age up to 50 is 26.1%, this number right here. Or we could say historically, 26.1% of Canadian Prime Ministers were in their late 40s are 50 years old. Now, the CDF is going to be the sum of the PDFs previously. And so we can use a trick that I showed you in an earlier video, we want to take the sum of the cell F 14, all the way to the cell F 14. And I could close the parentheses there. But I also want to add in this dollar sign to the beginning of the array, so that it doesn't change. And as I populate other cells with this formula, it's going to add all the PDFs up for us until we get to 100%. And there is our PDF and our CDF. Last but not

least, let's create a histogram using the Excel Analysis ToolPak. So the first thing you might need to do, if you're using Excel is to install the analysis tool pack. So I'm going to go to File, and I'm going to go to options. And I'm going to choose Add Ins. And if you don't already, select the analysis tool pack, and add it in. And if it's already added in, like mine's already added in, I don't need to do anything. So I'm just gonna press OK. And I want to go to data. And I want to choose the option here called data analysis tool. And when I do that, there's a whole bunch of different things that you can use tools that you can use. Lots of them are things that you'll learn in future statistics classes, but I'm just going to use the histogram option I'll press OK. And the histogram option wants two pieces of information once the data and that's the called the input range. So we could add the ages into the input range and it also wants to The bins sizes. And so we've got the bin sizes here. Of course, the bin sizes, you just need the highest value within that bin size. And Excel will figure out the rest. We want to chart output. And there are some other options here, we could have a cumulative percentage that sounds like a CDF. But let's just do a chart output. And let's see what that looks like. And we'll have it show up in a new, a new work worksheet, new worksheet. And you can add in the labels, all sorts of things. So we press OK. And here we are on a new sheet name labeled sheet one. And the analysis tool pack calculated all the frequencies for us, given the bin sizes, and then it created this kind of neat histogram. For us, that looks very kind of elegant. Now we might want to get rid of this little more category. So what we'll do is a highlight the more in the zero frequency here, and I'll clear those cells. So now the more is gone. Remember, there are no prime ministers who are inaugurated over the age of 75. And then if I click on the histogram, you notice that it's highlighted the bins and frequencies. And I can just click on the little corner there and pull that up. And now we've got rid of them more towards the right hand side. And we've got this nice histogram with frequencies. You can see Canada has had a lot of prime ministers that were inaugurated between 45 and 50. Most common age Prime Ministers get started, but we've had prime ministers that many other ages as

well.