

module_stat1_lecture13

Fri, 1/21 12:23PM 16:10

SUMMARY KEYWORDS

histogram, excel, bins, create, incomes, variables, data, visual representations, module, heights, column chart, borders, enabled, output, options, bars, tools, columns, generated, wanted

SPEAKERS

Sumon Majumdar

Sumon Majumdar 00:05

So the previous clip, I showed you how to create a histogram automatically in Excel. But one of the problems with that is that it's selected this bins, the starting point, the endpoint by itself, although you could change the size of the bins, the number of bins, but you didn't have as much control. So what if you wanted to create the pins on your own? How would you do that in Excel? So that's what I'm going to show you in the first bit of this clip. So we go back to our Excel file, right? So and this is the histogram we created, it started from 27 onwards, right? But suppose we wanted the bins to start at round numbers like 25 to 30, 30 to 35, and so on. Right? So how can we do that? So we can do that using a tool in Excel called data analysts. So the first thing, in most Excel programs, this is not automatically enabled. So what you need to do is go to the file. And, and at the bottom of it, you see, there's something called options. So you hit options. And if you go to the bottom of the screen, sorry, here, there is something on the left hand side, there is something called add-ins. So we need to add that. So at the bottom of that is what's called Excel add-ins. So let's do that. And now you see there are various add-ins which are possible. So what I've included is the analysis tool pack analysis tool pack vba. If you don't have those two check mark, then the analysis tool pack is not going to be included. So hit OK. And then this Data Analysis tab, it becomes active. So I've also noticed that sometimes you after enabling that, you may have to close and open your Excel just to make it enabled. But because the first time I did it, it didn't show up, Okay. So now what I want to do is I want to create my own histogram. So I want to create my bins. Right? So if I look at the heights, right, so if I look down, so what's the lowest height is 27. And if I look here, what's the highest one is 71. So I want to probably start from 25. Maybe I want to do this 25 the next- So the first one is up to 25. The second one is 25 to 35. The third one is 35 to 45, 45 to 55, and so on. So these are my bins. And now what I want to do is I want to create a histogram. So what I'm going to go do is go to data analysis. And see there is histogram, and I'm going to hit OK. Right? And what it asks you is what's the input range. So my input range is from here, sorry, let's go to input range. Clear what they have here. I'm going to start from here, that's B2. And where the where what's going to be the end of my input range is up to here, which is B31. Okay, so that's my input range. Now, what about my bin range? So bin range, let's go in here. So the bin range is going to start from C2 and up to where up to here, which is C77. Right? The other things that have to check off is the chart output. Right? And secondly, where do I want the output to be? Right? I can want it to be in a new worksheet, new workbook, or if I want to have it in the same sheet

here, come to my input range, and I hit somewhere where there is a free cell like this. And so that's where that's where my output is scraped off. And if I hit OK, now see what it has done is that it has created for me the bins and the different frequencies and created the histogram. Now see what it has done as compared to this histogram here, it has created more like a column chart, right with these as the different categories. But in the histogram, typically, these come next to each other with no gaps in between. So what I can do is I can go in and actually change that. So if I hit double click on any of these bars, right, so this format data series comes up, and see one of the options that comes here is the gap width, which determines the gap between the columns. So if you don't want that, right, so I'm going to change that gap width to zero. And now you see it starts looking more like a histogram with all the columns next to each other. So this gives me the histogram of, of my tree heights, right? So this is histogram of tree heights, and let's call it self generated. So problem here is that, since I got rid of the width, all of these bars, they're now they've run into each other, they're indistinguishable from each other. So how do I get them to put borders around them as there are in this particular histogram, so what I can do is if I double click, right, so a series of options come up. So if we go to the first one, which is the fill and line, now if I come to border, I can put in solid lines, and I can choose what color to put solid lines, if I choose black. Now see it has generated these borders. So this is how you can create your histogram using your own bins, right not depending on the automatic bins that were created by Excel. And again, you can change a lot of stuff, you can change the color, if you want to, again, replicate the same color as this, it's green, right? Now, again, you have to go in and change the borders. But you can do that, right. And you can also change the style, you can change the background. Again, Excel gives you a lot of options for doing. So let me quickly show you a one other example of a histogram again, so. So here is another sheet here in my same Excel file. This is a sheet of incomes, okay. So say you have different incomes in neighborhood or in a village, right. And what you're interested in knowing is what's the distribution of these incomes. So, again, you can either automatically get Excel to create an histogram, or you can create it yourself. So I'll let you do the automatic creation. So in, what I'm going to show you is how you can create it, create a histogram by yourself by choosing your own bin sizes. So suppose you want to choose your own bin sizes. And so what you may want to know is what's the maximum? And what's the minimum point of these incomes? Right? So an easy way to look at that is used the sort button in Excel, right? So if you saw this from smallest to largest. So just continue with the current selection sort. But it tells you that the minimum is at 1021. The highest is 4672. So what do you may want to do is create bins accordingly, maybe starting from 1000 1500 2000, and so on. So I've now created my bins for this. And I want to create a histogram by myself. So what I'm going to again go into is data analysis, go to histogram. Okay? And now I want to specify the input range. So it has the chosen B as the input range. I don't want that. I wanted to start from the two and where do I want to go up to, I wanted to go up to here, right? So I've chosen the input range. Now what about the bin range? Let's do bin range. So it has chosen C, automatically, I don't like that, I want the bin range to start from here and go up to here. And where do I want the output, so I want output somewhere around here. So I'm going to choose this. And I've chosen chart output. And now if I hit this, this should create a store. And what it also gives you is also generates the frequency distribution of each of the bins. So remember, what does this mean? It means it's those with incomes less than or equal to 1000. These are the people with incomes between greater than 1000, but less than or equal to 1500. The next bin is from top 1501 to 2000. The next one is 2001 to 2500. And there are six such there are nine in the previous been, and seven in the previous period. So right, so here is the histogram that it has created. Again, I don't like the fact that there are gaps between these columns, what I can do is I can go in and double click, and then change that gap point to zero. So then these are all next to each other. And I can go into here, put in borders, write a solid line with a block. So that's going to generate. So this is my histogram of, of incomes in this village. Right. So what we see from this

histogram is that there's a lot of people towards the bottom right, there are some people in the middle, but very few people at the top. So this is the type of thing that you want to use a histogram for to get an idea of the distribution of variables, whether it's in a village, or whether in a forest or anything else. So coming back, right, so what we have learned in this particular module, again, is to create a histogram in Excel, either automatically let Excel do it. Or if you want to have more control over the bins, that you can create it yourself. So this ends our module. So I'm going to very quickly summarize what we covered in this module. So we started with a basic introduction to statistics, the reason why statistics is important in the social sciences. And then I introduced you to the idea of where does data come from. And the important distinction between the population and the sample. So the population is the entire unit of study. And the sample is when you have a small part of the population. And when you have data on a small part of the population, you don't have the data for the entire population, but for a small part. Then we talked about the types of variables and there are two main types of variables. One is qualitative or categorical. So this is a variable, whether it comes in categories, the type of trees, the religion of a person, once opinion, like strongly agree, disagree, things like that. Or it could be a quantitative or a numeric variable with variable comes actually numbers, right, what's one's income, what's the height of a tree, and so on. And then we covered visual representations of data. So initially, we did visual representations of a categorical variable, which we did by using a bar or column chart, which could either be stacked if we wanted to show two data from two sources, or two sets of data, remember, like crimes, different types of crimes, but for different provinces? And the second thing that we did there is the pie chart, which shows you what's the percentages or what's the composition of, of different types of crimes within a given province. Now and then we did visual representations of numeric variables, and for that we use either a line plot to look at how trends change over time. Or if we wanted to look at the relationship between two variables, we use the scatterplot. And the last bit was a histogram, which we use to look at the distribution of the variable. So this concludes our first module on statistics. So the important thing here is for you to go into Excel and practice using the different visual tools that we learn, because those tools are super important to get an idea of how the data looks like. Or if you wanted to convey the idea not just to yourself, but to an audience of how the distribution of a data is or how the data is changing over time. These things like line plots or histograms, or scatter plots, or a bar chart, these are very important tools to have in your arsenal.