

module_stat1_lecture12

Fri, 1/21 12:23PM 9:05

SUMMARY KEYWORDS

histogram, bins, quantitative variable, distribution, trees, frequencies, plot, created, students, excel, chart, marks, interested, suppose, heights, data, different types, column chart, middle, scatterplot

SPEAKERS

Sumon Majumdar

Sumon Majumdar 00:05

So, so far for quantitative variables, we've talked about two visual tools. One was a plot for a time plot to show the trend of how cases or GDP is moving over time. And the second one was the scatterplot, which is to show you what may be the relationship between two variables. Now, there's a third aspect of quantitative variables, which could be important is the distribution. So for example, suppose you have marks in a class, see your 50, or 100 students in a class? And the professor may be interested in knowing how are the students doing? Are there a lot of students at the bottom half, or the top half? Or in the middle? or so on? So what we're interested in is the distribution, the same thing could arise. For example, if you want to look at the distribution of incomes in a country, is it very skewed towards the bottom towards the top? Or it's more weight or more mass towards the middle? So how do we look at how can we visually try and look at the distribution of a quantitative variable? So this is where our third two comes in important. And that's called a histogram. So what's the basic idea? So the basic idea is this that say, say you are interested in looking at the marks in class, the distribution of marks in class. And let's say this marks range from somewhere between somewhere from zero to 100. So what do you do is you create equal sized bins. So say you create a bin of zero to 10, then 11, to 20, 21 to 30, and so on. Right? And then look at how many students are in each bin. Right? So how many students have scores between zero and 10? How many have between 11 and 20? How many between 21, and 30, and so on, right? And then basically, so that's the frequency of each beam, and then basically plot the column chart of those, those of those frequencies, that gives you the histogram. So how do histograms look? So here is an example that I created using Excel. So this is an histogram where the data, so where the bins are of size five, so this is 27, to 32, 32 to 37, 37 to 42, and so on, right? And going up to the highest between 67 to 72. And what a plotted is the number of instances of this bin, so this is there are six, in this first bin there five, the frequencies five, for the second being, it's one for the third bin, to for the fourth bin, and so on. So this gives you the distribution. So looking at this, I haven't exactly told you what this is a distribution of that. But what it shows you is that there's a high frequency at the bottom and high frequency at the top, not so much in the middle. So this sort of an histogram can be created very easily by Excel basically by using one automatic. So Excel can create them automatically. Or if you don't like the way the bins are created by Excel, you can also create them by yourself. The main thing is that you want to create the bins which are of equal size. And secondly, you don't have want to have too many bins, or you don't

want to have too few. Because if you have too many bins, right because say in the marks example, you have your bins size of three by 10 to 13, 13-like 14 to 17 and so on, then you'll probably have one or two students in each of the bins. And it'll be very hard to analyze that sort of form a good idea from that that many bins, but if you have too few. So for example, suppose you have only two bins one is from 10 to 50 others from 50 to 100. That also doesn't give you a good idea of what the distribution is across the students. So now let me go in to excel and show you how you can create a histogram. So now I'm back to my Excel spreadsheet. And remember the example that we had done before of different types of trees in a forest, where it sampled a lot of 30 trees and written down the different types of trees. Now, suppose we add to that the height of each of the trees. So this is a numerical, quantitative variable. And what we may be interested in is the distribution of these heights. So what can we do? So let me select the data from here. And now I go to insert and see among the various things here, one of them is this insert statistic chart. If you hit that, the first thing that comes up is the histogram. So, and here, Excel has automatically created this histogram for you. And what it is done is that it has looked at the data. And the minimum point on the data is if you go down, it's 27. Right. So that started the first bin started from 27. And somehow, it has automatically created three bins and generated the frequencies and plotted the system. Now I don't quite like that, it has only created three bins, right, so probably want a little bit more. So what I can do is I can change that in Excel. So if I go to the bottom and hit double click there, right. So see, format axis comes up on the right, and their axis options, and one of the things is bins. So you can choose the number of bins that you want, you can also choose the bin width. So suppose you want to choose bin width of, let's say, five. So for hit that, right, so it generates now quite a number of bins, each of them have width five. So I like this distribution, because what it shows me is that most of the trees are either short on the short side, or on the tall side, there are very few trees in the middle. Now as as with other charts, here also, you can do a lot of different types of customization, you can change the color, if you want to make it green to reflect trees, you can do that you can choose different styles. So something like this, you can put in a chart title. So this is the histogram of tree heights. So you can do lots of different things with this histogram, as you could do with all the charts. Now, there may be one particular thing here that you may not like, again, the type of type of bins that it has automatically created, created from 27 to 32, 32 to 37, 37 to 42, and so on. Right? So maybe you want holds numbers, right? So you probably want to 25 to 30, 30 to 35 and so on. If you want to create your own bins. There is also a way to do that in Excel. And that's what I'm going to show you in the next