Research for Marketing Decisions

Vaibhav Chawla

Department of Management Studies

Indian Institute of Technology Madras

Week - 07

Lecture - 33

Data Analysis using SPSS: Charts, Outliers, and Descriptive Statistics

So in the last class, we finished the data cleaning, there are few more steps left that we will be taking up today, and then some some form of descriptive statistics, and then inferential statistics, as well so once you collect the data you enter the data in this form, the sheet that you see. If you have a sheet for you to help prepare your SPSS data entry, variable and data view, make a sheet like this, do the data entry and begin the data cleaning as it is taught to you in the last class. Whether it is data related to human resources, strategic management, employees data, customer data, even with the financial data also you will have to do the same thing. Without data cleaning if you proceed the results could be erroneous, the result could be inaccurate, right? so data cleaning is the basic thing that one should know so what all different data cleaning techniques we discussed in the last class, we begin with looking at the range,

maximum, minimum so that we get to know whether there is any value that is marked outside the range and we catch it and we clean it, we correct it. Then missing values and then correcting them. Right? Today, we are also going to look at how to catch the outliers or extreme outliers and what to do with them?

so we will that is also part of data cleaning, so let us proceed with the step number four, and we will soon come to the outlier analysis as well, step number four is construct a frequency distribution of the variables, gender familiarity with the internet, whether than shopping on internet, whether than banking on internet. It's very simple, one has to go to analyze descriptive frequencies, what all variables are there, gender, familiarity with the internet. Whether done shopping on internet. And whether done banking on internet. And display frequency tables.

If you want just keep that all these four variables into the operations box and construct the frequency distribution so this which means, take the display frequency tables you will be able to see the frequency table for each of these variables that will give you an idea where more people are marking. Now you will get an idea of the mode mode, which is a central tendency, which has the largest frequency which value right?. Now beginning now going to the step number five, which is, so here we are just using this different descriptive statistics to to know our data in terms of we will be looking at mean, median, mode, more here with the frequency distribution we are going to look at the mode, and it is right in front of you uh. The next one is we will make a chart frequency chart not distribution, the fifth one if you look at the fifth one is, make a make a chart for gender usage of internet shopping usage of internet banking. Now what are these kind of variables pie chart you make for what kind of variables? categorical variable, nominal and categorical, those that are nominal and categorical you make pie chart. If you have to show the frequency graphically to somebody for nominal variable which is also categorical, use the pie chart.

Graph graphically it is easy to understand things right? that is why with the frequency distribution the table it is little we have to read the numbers, but when we look at the chart it is more simpler for us to understand right because our probably our visual uh visually we are made graphically we are able to make more sense of things quickly than by looking at the data. So the fifth one is we have to make a pie chart for these three variables, now go to there are two ways one is grow to go to graphs, and let's go to chart builder, click on chart builder, and click OK, you will get a screen like this those this one is about graphically representing the frequency. Now there are three kind of graphs pie charts, bar charts, and histogram, depending upon the type of variable we choose we should choose the appropriate graph. If it is a nominal and a cat nominal variable and categorical because just nominal doesn't make a nominal means with your names doesn't make any sense nominal and categorical one should always use pie chart other graphs do not make sense. Bar chart you can use for interval histograms for continuous data so here we are talking about the idea of putting this fifth point was to let you know that what kind of charts are suitable for what kind of variables, so let's now draw the graphs for pie chart for these three variables.

Go to graphs chart builder. Click OK and you have a screen like this. Now you want to plot a pie chart. So pie chart comes drag it here and drop it. Once you drag it, drop it, you have some element properties also open just close this one and we had three variables

including gender, so first let's draw for gender if you click OK you are you are able to see that both are equal 50 50 percent of whatever data numbers are there.

How you can know what is the frequency of male and female, just double click it, and you will get a screen like this chart editor just click this show data labels, and here displayed is the percentage. If you want to display count, move it here and click apply and you will also be able to see. I will do it again. No worries. So I am again going to draw the pie chart for gender variable.

Analyze, oh sorry, graphs, chart builder. Okay, you have this. Here you look at okay then this screen chart builder screen will come, then wait wait wait then choose from PI and polar here and drag once you Pie and polar this Pie chart comes here now drop it, element property you can close at this point of time, right? now if you whatever variable you would want to do it, let's say, you want to do it for gender drop it here, and then click okay and you are getting this, now you want to know what is the frequency for each gender left double click, and you get this here the icon says show data labels click on this, and what is getting displayed in the percentage, move the count also so that you know the count and you will apply then you will also see the count.

If you do not want to see the percentage, you can bring it down. You can actually remove it and apply. So, percentage will go. And if you want to move the label, let us say like this. So, you can do that also.

Outside. You can change the color. You can play with it. Now do it for other two variables quickly, so let me do it for next one which is banking on internet banking on internet reset, this one here close it bank shopping on internet okay this is there double click show data labels move the count up click apply and if you want to custom this is what you get for a third one also quickly do it SPSS is working for everybody.

Slightly different pie chart. Slightly different pie chart yeah. okay i'll plot for the other one last one as well quickly graphs chart builder, okay, reset move it there, cancel it, whether done banking on internet here okay you got this chart double click show data labels move the count up apply, you want this to be displayed outside, choose custom, select this apply and close and close and you get sorry okay, no no but it is okay okay I did it for three, okay, I will do it really slow for okay graphs chart builder let I am doing again for gender, okay, I'm resetting, so from choose from you select pie pie chart comes, drag it here element properties close at this point of time you want for gender, move

gender to slice by, and click OK, you will be able to see a colorful division of a nominal and categorical variable if you want more left double click you will get a screen like this

and this icon show data labels click on this properties will be activated, in this property screen will be activated here percentage is already displayed just because we activated the data labels if you also want to see the count click on count which is not displayed move it up and just apply and close it that's it we don't do anything more, that's it so now we move to the next one, construct the frequency, the sixth one construct the frequency distribution of the variables internet usage in hours per week, attitude towards internet and attitude to our technology, so you have to tell me what you will draw for which variable Internet usage. What is the kind of variable? So, which means for internet usage in hours per week, it is a ratio and other two.

You have to tell me other two also. Attitude towards internet and attitude towards technology. You are caught today. You have to tell me. Attitude towards technology.

What is the kind of variable? Ratio. Again. Okay. So, attitude towards internet and attitude towards technology.

What is the kind of variable? Interval. So, what we will draw for them? Bar chart for IUPW internet users in hours per week. Histogram.

Please do that. So we have to go again to graphs. Chart builder. Okay. Reset.

Simple histogram. Drag drop. Close it. IUPW. Right?

Move it to x-axis. Click okay. So, it doesn't look like a normal distribution right? for this particular sample you got it everybody has been able to draw it it's very simple now, once you know where it is the chart builder you don't require me also to now so, the next one is we have to draw bar chart. Bar, let's reset it bar chart drag drop so attitude towards technology and attitude towards internet right attitude toward internet click ok we have this and attitude toward technology ah in the sample the So, we have done the step number 6 as well.

This is more easily, easy to look than the frequency table. Just the purpose is to let you see it. It is easy to make sense of it graphically than with numbers, right? Here also numbers are there, but it is little easy to do this. Histogram as I said for continuous variables IUPW internet usage in hours per week bar chart for interval right now the next

so, if your manager tomorrow says you have data and analyze and graphically show you know what to do what not to do that is more important right the next one is 7th point,

where we will do only the box and whisker plot box and whisker plot all of you know box and whisker plot so box very good so box and first quartile second quartile third quartile box and whisker plot we use to identify outliers in the data and extreme outliers so it is also part of data cleaning so which means if say if if let's say You are asked to determine the, you are asked to find out the average income of the senior batch just graduated. Right? And you see that, let's say class size was 50. You see that 59 have the average, 59 have the salary between 20 to 40.

And there was one who has the salary of 3 crores. Let's say, now, if you calculate the mean, it will be biased because the 3 crore salary person he is pulling the mean higher, whereas this mean will not be representative of the class because most of the class has the salary in between 20 to 40 and this person 3 crores he is pulling the mean towards upside. So, and if we do not identify this outlier and just calculate the mean it will be incorrect central tendency to show right? So, in order to identify any outlier in the data we use the box and whisker plot.

So box and whisker, let's look at box and whisker plot. Let's say I am writing the marks of 9 students in market research at the end of the course that's the seven students right? so the first I want to identify whether there is an outlier, now the first thing is I would arrange them in either ascending or I would arrange them in ascending order so starting from the least one 10 30 50 60 70 80 90 or let me change this 30 to 75 so 10 50 60 70 75 80 90 so ascending order after that I identify the median this is the median central the the middle value, if this if let's say, the data points are 8 so after any ascending order I would look at that middle 2 value and take their mean to identify the median, so this is the median so median is 70 this is second quartile median is the second quartile now what I will do is

I will look at the first quartile, and the third quartile, so first quartile is 50, third quartile which is central value of the values on the left will give me first quartile, this is quartile 1, this is quartile 3, so I will draw the box with quartile 3 is 80. And quartile 1 is 50, so box is ready, now whisker whiskers start from here and go till what point 1.5 times interquartile range which is what is interquartile range, interquartile range is quarter quartile 3 minus quartile 1 which is 80 minus 50 is equal to 30. So 1.5 times 30 is 45 so Q 3 80 it cannot go higher than 100 so if I add 80 plus 45 it'll become 125, which is not

possible, so the in this case the maximum risker will be till 100, let's say, this is 100 but it cannot because it cannot go beyond on the left side again 45 which means at 5 right so any value which is outside 5 less than 5. Greater than 100 is not possible.

Less than 5. Let's say if there is one person who scored less than 5. So that will become an outlier. And once you identify an outlier, you have choice. If it is the right response, if it is not, sometimes outlier can be because somebody did not pay attention and just filled their response something.

Wrong response. Biased response. But at other times it could be the original value, if it is original let's say, in this case if it is original then I should not represent the class marks through mean because the marks this outlier is pulling the mean towards the downside, I should rather use median, the other thing is sometimes in analysis you remove the outlier respondents, who have given outlier responses to many variables, you remove them so that that is also part of data cleaning.

Sometimes you remove these let us say this was the this was the respondent who had three marks which is let us say it is not here, of this was a person who had three marks, or when you are doing a market research survey there is a respondent who has given three or four outliers when you look at the variables so that person is a candidate for removal from the data that will make your life easy when you are doing data analysis, so that is how you identify the you identify the outlier. Now we are doing the seventh one, where we are using the explore feature to make box-and-whisker for I UPW internet users so what to do go to analyze descriptors explore internet users in hours per week we will move it to dependent list and click on plots you already have box slot let's not stem and leaf if you want I will show it but I will redo it, so go to analyze descriptive statistics, explore whatever variable you want to do box plot for, move it to dependent list go to plots, in plots already box plot is selected by god, and then stem and leaf is also selected let it be there, we will not look at stem and leaf, not required really, it just tell you stem and leaf tells you the shape of the sample for a particular variable whether it is negatively skewed, positively skewed, or normal. If if the shape of the sample is positively skewed,

this is called positively skewed. It is positive toward the larger number side. So let's say the class marks in market research. This is 50, 55, 65, 75, 85, 90. This is a positively skewed.

In positively skewed, mean will be in the front, median in the center always, and mode is here. Whereas in negatively skewed the order will again mean will be there as the

Mahamantri the prime minister right in front, then median and then mode ok, so yeah we were doing the box plot. So let me do it again quickly. Analyze descriptive statistics, explore internet usage, dependent list, go to plots, already selected, continue. You can look at what is there in statistics, nothing else, nothing more, nothing much, click okay.

You want statistics and plots, plots, okay. Both are okay, give us both. So this is the box plot. This is this is the one we are getting, if you want to activated double click, and you want to again let's not do that for this, so this is the box plot, let's say, we want to look at so here we are getting quartile 1, quartile 2, and median right? for this. Let's say, you want to see whether there is a difference. Now this is there is no outlier, do you see any outlier when the outlier will be there there will be a dot, there will be some dot here, I am not able to draw it will be some draw dot or star, star means extreme outlier, okay, but box plot is important for another reason.

So let me show you that we can compare for two genders whether the internet usage in hours per week graphically whether it differs. So how to do that? Go to chart builder. We are doing box plot only. Box plot.

Move it here. And we want to see it for gender and I UPW and click OK, so you see between two genders whether their internet usage in hours per week, what is the distribution dispersion in the sample, so you see you will this is another good way of showing for the nominal for a particular categorical variable, whether they differ on certain variable so you can use this just. Graphically it is easy to see.