Research for Marketing Decisions

Vaibhav Chawla

Department of Management Studies

Indian Institute of Technology Madras

Week - 07

Lecture - 34

Data Analysis using SPSS: Cross Tabs

Let's move to the next one where 8th one, where we want to look at the compute the descriptive statistics for internet usage, attitude towards internet, and attitude towards technology, so, descriptive statistics we want to look at mean median mode standard deviation and range skewness kurtosis and so on so because we have now corrected everything cleaned the data, now we are looking at descriptives with the outlier the data cleaning ends, then we look at the descriptors, descriptive statistics, so descriptive statistics analyze descriptive statistics for what all we need to look at internet usage attitude towards internet, attitude towards technology, options mean is there, standard deviation, minimum, maximum, variance, range, contours, skewness, ok, click ok, you will get it. Sir, can you explain like what that this graph actually interpret from the. Male and female what is the median internet usage in hours per week? what is their quartile 1? what is their quartile 3? what is the quartile 3 for males? what is the quartile 3 for females? so you can compare the distribution, what is the, so here you can also see the maximum value for females, there is no whisker all itself so you can just compare that the values are this spread

on the more positive side are the higher for male versus female. You can compare their internet usage graphically using this. Okay, 1.5 times interquartile range. Yeah, this is the length of the whisker 1.5 times IQR it is the length of this whisker, any number which is beyond this in positive side, it is not possible, because maximum marks are 100, and if I add 1.5 times IQR IQR is 30 it become 45, so 80 plus 45 125, 125 not possible because the question is the answer question paper is out of 100 the exam marks are from 100 but in negative side, it is possible yeah 1.5 times IQR on the positive side and negative side on the negative side yeah we go again 1.5 times IQR here so 50 minus here it was 50

minus 1.5 times IQR was 5 this you will have to do for your project everything it is very easy I think you would have already done many of these things right

in Excel you would have done now the next one ninth one is now we have under we have cleaned the data we know the descriptives mean median mode range mode for categorical variable mean for interval and ratio variable. Mean for interval variable, if correctly it should not make sense but it is used so because it is there is no absolute zero so when you take the mean but it is used in the analysis. But arithmetic mean we can use for intervals. Arithmetic mean we can use for interval scale. So moving further.

Ninth one. Do a cross tabulation between gender and internet usage. Now we are going to look at once. So this is the problem for internet use internet service provider. He wants to know what determines internet usage in a what determines the internet usage which variable explain it.

So our role our work does not end here. We want to look at which variables explain internet usage in hours per week. What is which variable have the relationship if male uses more I should target male population. I should spend money on them so that they use it more, I go give offer to them change my offering based upon how they use it right? so I want to look at the relationships now between the dependent and independent variables so the simple way is by bivariate, this is bivariate analysis we are doing the simplest technique is cross tabulation what do we do in cross tabulation? let us say in covid times we want to look at in covid times, we want to see whether there is a whether there is some pattern between or whether there is some association graphically we want to see whether

there is some association between gender, either male or female, and the stress they felt during COVID times. So let's say I have about this. These are two variables. Male. First respondent is male.

Stress is high. Second is female. Stress is low. Third is male. Stress is low.

Fourth is male. Female stress is high, fifth is male stress is high, sixth is male stress is high, seventh is male stress is high, eighth is female stress is low, ninth is female stress is low. Let's say it goes till 30 I want to just draw a cross tab, which means I want to cross these two variables categorical variables to see whether there is graphically can I see there is some association, whether male or female who is feeling more stressed who was more stressed during covid times, so what I do here gender variable male female cross-

frequency table it is cross-tabulation and stress high low i did for 30 respondents so 25 males had high stress and 5 low, 5 females had high stress and 25 this is simple cross-tabulation using this, one can see whether there is an association graphically not statistically but graphically whether there is an association between two variables or whether there is a frequency distribution of or not. Difference in frequencies with respect to a particular variable for two genders.

So cross tabulation. This is easy you can see. Gender. So. Generally we.

What is the independent variable here? Everybody. Obviously because gender stress maybe or may not be determined by gender, but gender doesn't get determined by stress right so we write independent variable along the columns and dependent along the rows, so this is the cross tabulation now we need to do the cross tabulation for gender and internet usage can we do that can we do that?

Can we do the cross simulation now? Internet usage at what variable? Ratio. What I said chi-square? With categorical, two categorical variables.

But our internet usage in hours per week is ratio sorry internet usage in hours per week is not categorical categorical is are generally nominal variables in which have either two categories three categories you can so internet usage in hours per week is a continuous variable it's a ratio variable so how can we use crosstab Can we convert so as I told you we can convert a ratio variable into categorical but a categorical into ratio we cannot convert. So what to do how to convert IUPW to a categorical variable so that we can use the cross tabulation. One way is we can divide, we can ask the internet service provider what do you consider as the midpoint between high and low IUPW, internet usage.

Other ways you look at the mean for internet usage hours per week, mean is 6.6 here. Otherwise, you divide on the mean, less than mean, low internet usage. Higher than mean, high internet usage. Let's say internet service provider as it is given in your book also. Internet service provider says 5 is the value I want to take cut off between high and low.

Ideally, one should proceed for mean or median. But let's say in this case, let's go with 5 you can see the same in same results in your textbooks as well, so if we take five is the cutoff between high and low we need to create one more variable IUPW categorical, how to do that? go to transform, recode into different variables, click on this, we want to

transform internet usage in hours per week to what new variable? Let's say IUPW underscore categorical.

And let's give that the name and the label. Let's keep it same for the saving time. Once you do that, click change. So go to transform, recode into different variables. Internet usage in hours per week.

We want to convert it to categorical. Move it to the operations box. What is the new variable we want to create? IUPW categorical underscore C. Let's say we write categorical. Same name and label.

Let's keep the same like some people have the same name and the small same nickname also, name and label they say somebody's name is I know it's so short you cannot I know I know it is so other people would have big names and then you can keep there so it's like name and label we are keeping same here click change and now we have to click old and new value we want to tell the SPSS that we want to create the new variable IUPW underscore categorical and how to make it categorical click old and new values here range lowest through value which means equal to or less than 5 equal to or less than 5 let's give value 1 which means low internet usage add 5.001 range value through highest which means greater than equal to this value lowest was less than equal to this value so we say value 2 click on add and click continue click OK you have new variable created which is categorical iupw underscore categorical have you done ok let me remove it go to transform compute

sorry transform recode into transform recode into different variables internet usage in hours per week I want to make it IUPW categorical or let me ask keep the same name as same label for this name categorical I click on change so then which means I am going to transform iupw to iupw underscore categorical but still now I did not tell SPSS how to convert the continuous variable into categorical that I will do clicking on old and new values in old and new values click on this range lowest through value which means some number less than equal to that number so if internet service provider has told that let's choose 5 5 or less than 5 low internet usage above 5 is high internet usage how I should convert IUPW into that categorical variable range lowest through value is 5 or less than 5 how I will tell SPSS 5 or less than 5 and give its value of 1 and click add now what about values greater than 5 so I will say 5.001 and greater than it equal to 5.001 or greater than it gives a value of 2 click add click continue and click ok and you have the new variable and new categorical variable created Did you get it?

Now how to do the cross tab? We have to go to analyze. Descriptives. And there is cross tabs. Click on the cross tabs.

We are getting this particular table. Columns are the independent variables. So we want to do between gender and internet usage. So where gender will go? Rows or columns?

Columns. Gender is a dependent variable or independent variable here? Dependent. Independent. Internet usage will determine gender or gender will determine internet usage?

Sorry? Here we are not looking at cause and effect. We are just looking at whether they are associated. We are not saying one is a determinant of other, one causes other. No.

We are just saying there is an association. But if common sense it says that gender cannot be determined by. So let us keep it as if at all there is a causal, if at all there is some association and there would be some cause and effect if at all. Then gender should be on the column side. So let us keep gender on the column.

IUPW categorical on the row side, click on no statistic, no don't click click on cells, and we want to calculate the percentage column wise in the direction of the independent variable which means you will get to know how many percent of males have high and how many percent have low like for females in the direction of the independent variable if you click ok you are getting this are you getting this once more I take extra payment for once more ok analyze descriptive statistics cross tabs, let's reset so gender which is the which could be the independent which is the independent but may not be causal but let's it makes sense to put gender as very independent variable in the column IUPCW in the row side click on the cells and we will calculate the percentages across or in the direction of the independent variable which is columns click continue click okay you will see this which means 33.3 percent one-third of the males have low internet usage and two-thirds have high whereas this order is reverse for the females now let's go to the next one where we are doing a multiple cross tab now what is this animal

If you remember, we discussed an example in a particular class where we were looking at the cross tabulation between two variables. One was education and owning an expensive automobile. Do you remember? The percentages were showing that higher education, people with higher education, they own the expensive automobile. The cross tab was showing that.

But then we thought, does education actually determine who will have the most expensive car? So we brought income as the control variable. And when we brought income for high income and low income, we drew the cross tab again to look at whether there is some, whether education determines Whether people will own expensive automobile or income. So when we brought income, that relationship of education and owning an expensive automobile went away.

So which means when we bring a third variable, when the original association goes away, that the original relationship goes away that we call as spurious relationship, so in order to if we have certain doubt whether that if we the association between gender and internet usage may not be because of so gender may not be leading to, so gender may not lead to high or low internet usage maybe gender has some effect with something else which there could be some other variable now in this case that could be familiarity with the internet, so we will now do a familiar cross tab bringing in familiarity with the internet as a control variable to see whether that relationship still association like this still exists, so what we do now is cross tab between three variables using familiarity with the internet as a control variable now familiarity with the internet is what variable interval It is not categorical. So we have to convert it to categorical.

What is the midpoint? 7 point scale, 4 is the midpoint. Can we divide it 4 or less than 4, low familiarity and higher than 4, high familiarity. Can we do that? Then do it please quickly

familiarity with the internet underscore categorical click on change hold a new values 4 or less than 4 value 1 and since there is no four point one, four point two, four point three, because it is a discrete interval variable which has values one to seven so we can say five or above value, oh sorry first we did not complete that four and below value one five and above value two and click add and click continue and click ok you have a new variable created familiarity with the internet categorical, now once you create these categorical variables in the variable view you can actually update these values also, IUPW value 1 low IUPW, value 2 high IUPW, same way we have to do it for familiarity also value 1 low familiarity with the internet, value two high familiarity with the internet, so we have now familiarity also as a categorical variable, so now we will do then multiple cross step tenth, so how to do that? go to analyze, descriptives, explore don't go to explore you tell me I'm just checking whether you are paying attention, go to analyze, descriptives, cross tabs reset familiarity now control variable layer we will add that in so gender so IUPW row

gender column and familiarity as the control variable or the third variable that we will introduce to see whether still the association between gender and IUPW holds, so again go to statistics or not statistics sorry cells and click on percentages row and click ok and you will get this, and I will I am not going to interpret because you interpret on your own please. So for low familiarity, you will have some association. For high familiarity, you might be able to see some association. So the idea is this is a way to do multiple crosstab. To introduce a third variable to see whether the relationship between original two categorical variable is still as clear, still holds, that is not a spurious relationship.

That we will only get to know whether familiarity will better differentiate between high and low internet usage when you do a crosstab. Here we are seeing whether that relationship between gender and internet usage in hours per week whether this is spurious or not. So what is the next one, now cross tab will give us a graphical examination, cross tab cross tab will give us a graphical way of looking at the Association, for example when we look at this data this between gender and internet usage, what so we were able to see that see that one-third of the males have low internet usage whereas two-third have high and it is a reverse for females, so what would you temp what you would be tempted to infer saying that you would say that in the sample it looks like females have males have more internet usage in hours per week than females right? if

so this is with the graphically you would be able to say male they have a higher they use they have the higher internet usage, because the frequency says so 66.7% as compared to 33.3%. Is there a statistical way to say that the difference is there? This is through looking at the crossfab but there is also a statistical test that is called a chi-square test, so which means we have now come to now this shows in the sample whether you'll see there is an association in the sample you are able to see whether so you are able to see males they have more higher internet usage in hours per week than females, but this is for the sample, we don't know for the population, so now we will move one step further to make inference about the population from the sample for definite findings so that one could actually use them in their business problem. So now we have come to hypothesis testing where we will begin with the where we will begin with the chi square.