Research for Marketing Decisions

Vaibhav Chawla

Department of Management Studies

Indian Institute of Technology Madras

Week - 07

Lecture - 35

Data Analysis using SPSS: Hypothesis Testing for Association

So hypothesis testing is two types, so there is test of association, and test of differences. We are going to start with chi-square, which is a test of association. So whenever we are doing hypothesis testing we are using a sample to draw inferences about the population, from the sample we draw inferences about the population to which the sample belongs, so now we are going to use the chi-square test for this cross tabulation saying that from the sample can we say something about the population that in the population there is a strong association, significant association between males, between gender and internet usage in hours per week such that one gender is significantly using higher rate than the other. Can we statistically say that about the population? Because with sample, you might have chosen a sample that is convenient to you.

If it is was not selected random, and it is too small of a sample 30 so only once you make inferences about the population then you could move aiming that you will target it for males because they have significantly more internet usage in hours per week, so we move to the chi-square test which is the first test of association between the categorical variables to see whether there is an association or not. So anyways so coming to the chi-square test so the null hypothesis and alternate hypothesis, null hypothesis in chi-square case is that there is no association, alternate hypothesis there is association, and we will test this now using the chi-square method. Now chi-square we look at the critical value and this is the 5% zone and what it means we will come, we will see in a while once we do the chi square test so lets say, in this case there are you know we have male female gender male female and IUPW high low we had..... 5, 5, 10, is it something like that right? now this is a sample this is in the sample let's say if what is the null hypothesis that

there is no association, if there is no association the frequency there would have been equal so which means total is 15 15 15 15 25 sorry here total is

here total both sides is 30, so if we calculate the when there is I think we multiply, so in the null hypothesis we say that there is no association when there is no association the frequency in all the four should be equal which means I think it is calculated by you multiply the for the particular cell you multiply the column total by the row total and divide by the total number of respondents so which means 15 into 15 divided by 30 will be how much 7.5 225 divided by 30 7.5 so this is the this should be the frequency actually if you calculate this will be the frequency in all the four cells this says if there is no association if the null hypothesis is true then the case should be 7.5 should be frequency in each cell right? you multiply for that but if you want to calculate the so null hypothesis in this case is, there is no association, alternate is, there is association. No association means the frequencies in each cell is same. So if you look at the null hypothesis, there is no association.

So in hypothesis testing, we start with the assumption that assuming that null hypothesis is true. If we assume that null hypothesis is true, there is no association. What would have been the frequency in each 4 cells? 7.5, 7.5, 7.75. Right? That is the expected frequency if there is no association.

If null hypothesis would have been true, what is the frequency? What should have been the frequency? That is called expected frequency. 7.5 in each cell. And there is something called observed.

We got from the sample which is 10, 5, 5, 10. So if the null hypothesis is true, Which means there is no association. The difference between the observed and expected should be less than the critical value. There is some critical value.

We will come to that. What is that? So what we do is we calculate chi square with summation of observed minus expected square divided by expected. And we will get some chi square value. If we do observed minus expected for each cell divided by expected, here you are getting something called 3.33.

And then we look at the chi-square chart for alpha level 95% and degrees of freedom. How we calculate degrees of freedom? Row minus number of categories in the row minus 1. 2 minus 1 into 2 minus 1. Both sides is equal to 1.

So what is happening here is, we have a sample, and we are calculating a chi-square value and looking at some critical value and trying to infer for the population that in the population there is some association or not. as I told in the hypothesis testing we assume that null hypothesis is true if the null hypothesis would have been true the difference between observed and expected would have been zero. Now we are calculating how much is the difference and if the difference is so huge so huge what is that huge critical value? if it is larger than that then we say that it is not because if the difference is larger than the critical value then we reject the null hypothesis so the idea is we say that if really in the population there is no difference if really in the population there is no difference there is a huge chance that you would have picked up the sample where the chi-square value would have been less than the critical value. This critical value is chosen.

This is the 5%. This is the 95%. Which means we are saying that if in the population there is no difference between the, there is no association. Then there was, there is a huge chance that you would have picked up a sample whose difference with the expected is not much but you picked up a sample whose difference was more than critical value if you pick up such a sample then which means your initial assumption itself you are assuming that in the population there is no association and then when you assume something you are expected to also draw a sample in which there will be no association but you

drew a sample in which there was huge difference and if that happens then you conclude that my initial assumption was wrong because in this sample there is significant there is a huge difference which means in the my assumption H0 null hypothesis I should not accept it reject the null hypothesis I will tell you about the this is what is called P value we calculate the probability and we look at the chance what is the chance we will so here it is the 5% critical value is the 5% zone probability that we would obtain such a sample which has such a large probability to obtain a sample from a population where we assume there is no association the probability of getting a sample from the there is a 95 percent probability that from the sample where there is from the population there is no association you will get a sample where there is no association and there is only 5 percent chance that you will get so when you are drawing if you get a sample which has significant difference With the expected, you conclude that you reject the null hypothesis. We will see, we will take it in a more clear way when we move to the sampling distribution in the T test.

It might become little more clear if there is some confusion here. So do you understand the idea of p-value and critical value? So simple is we assume that H0 is correct. If H0 is

correct which means in the population there is no association. Now if in that population you draw a sample it is like if your null hypothesis here assumption is right in that sample the observed minus expected difference should have been less.

But you are getting a big difference and that is big difference such a big difference that you are likely to obtain that difference only by only 5% of samples from that population, and you cannot be so unlucky that you pick up that 5%. So as you assume that. You assume that you are not so unlucky and this there is a significant difference.

Your assumption is wrong. So you conclude that you reject the null hypothesis. So, we will do the chi-square test now and go to analyze. It is the same. Go to the cross tabs and chi-square.

We have to go to the cross tabs. Just we have to now click on the statistics and analyze. What is the difference between critical value and free value? Critical value is the chi-square critical value, which means the difference. How much should have been the difference?

So, it is it. So, let me so critical value it will give you cut off critical value is the value which you obtain by from this formula and it will be in some units units of or it will be the it is a frequency so you will get some value that this higher is the difference between the observed minus expected now how high is too high that we reject the null hypothesis that is the critical value which decides if value is larger than this critical value then we say that so higher than the critical value there is a critical value higher than the critical value the probability value is the probability that you will obtain a sample that you will select a sample that has a difference chi-square value of greater than this critical value that is a probability how much probability is there so alpha that we say alpha is 0.05 which means five percent probability which is PP is equal to 0.05 so we say greater than let's say if this critical value is 3.1 which means if how what is the probability

that you obtain a sample higher than this difference and this probability here it is written only 5 percent probability that you obtain a sample higher than the you obtain a sample which has chi-square value higher than 3.1 so we go to analyze descriptive statistics cross tabs there is a video also I made a video I will send you the link for the chi-square so let's do this first chi-square analysis so we have rows 9th one we are doing 11th one do a chi-square test between gender and internet user gender is row and column is IUPW column cells this one and in addition to this now we will calculate the a test statistics chi-square statistics and we will look at the probability value p value if the p value which is the

probability value is less than 0.0 less than equal to 0.05 which means in the critical zone zone then we say that then we reject the null hypothesis otherwise we will not reject the null hypothesis oh sorry column yeah sorry column should be gender yes so descriptives crosstabs reset column is gender IUPW here statistics we will put chi-square and click on this is fine chi-square is fine in cells percentages So in the statistics you have chi-square and other measure also like phi coefficient is there, contingency coefficient is there.

They are also useful but we will stick to this chi-square. They are also chi-square statistics but little adjusted. So if we click OK, we are getting 0.068. this is two-sided we have to look at two-sided this one asymptotic significant probability value is 0.068 which is greater than 0.05 and when we are using two cross two matrix it is a special case so we do not look at the Pearson chi-square but we look at the continuity correction for 2 cross 2 and for 2 cross 2 or when the observed frequency in any cell is less than 10 we use the continuity correction let's say if this would have been 3 cross 2 and observed frequency in any cell every cell is greater than 10 then we need not use this continuity correction

we have to look at the two sided p value so p value is given 0.068 but we are not going to look at that because this is special case 2 cross 2 so we are going to look at 0.144 this value continuity correction it is an adjustment to the chi square value and this we look at in special case when we are looking at 2 cross 2 matrix or in any of the cells observed value frequency is less than 10 and we should use chi square only when the observed in each cell is greater than 5. Otherwise, we should not use it. Now, do a chi-square for familiarity and internet usage, which is the last one, 11th one. So, we conclude, so p-value if you look at 0.144.

Which means it is not greater than the critical value which is 3.33. 0.144 is the probability value which is greater than 0.05. It is only when the probability is less than equal to 0.05 we conclude that there is an association. See, it is very simple. So, when somebody, let's say, in a murder case, when somebody is an accused, okay, in a murder case, when somebody is an accused, so court, what does it say?

Innocent until proven guilty, right? Innocent until proven guilty. So, null hypothesis is innocent. So, here also, no association, no difference. That is what is null hypothesis.

And court says that we assume the person is innocent until you have a strong evidence here also what we are saying we assume there is no association until you have an evidence so we start with the assumption that we assume null hypothesis is true let's look at any evidence contrary to that that is what we are doing in hypothesis testing we follow

the justice process and say that there is no association there is no difference then we look at any evidence against that assumption. Sir, continuity correction, it is given to two-sided tests. Sir, and can you please explain again what exactly it is done, what it does?

So, continuity correction, it is some, it is not the mathematical formulation, it is not within the scope of this course, but it corrects the Pearson chi-square which is which has little bit of incorrectness in case of 2 cross 2 matrix some adjustment is required to improve that and that is why we look at the continuity correction in the special cases when it is a 2 cross 2 cross-tabulation or when in any of the cell the the observed frequency is less than 10 then we use continuity correction the p-value corresponding to continuity correction rather than looking at the Pearson chi-square. Yes, we look at two-sided p-value here although because the chi in the chi-square we we are looking at in the chi-square we are looking at only one end, and we are giving total ah generally when you see if you are looking at two-sided so it will be point zero two five one side point zero two five one side but in chi-square we look at point zero five on the yeah but here, the value which we are going to look at is under the column of two-sided, whatever value comes here. Asymptotic significance, we are going to look at the value under the two-sided.

But it is the value which is on the further side of the chi-square, which is a positively skewed one, and it cannot be negative, because we what we are doing is observed minus expected square, so we it is made in such a way that it will have positive values only. Do the next one between familiarity and internet usage, now familiarity and internet usage both are categorical go to analyze, the next one is 12th, do chi-square between familiarity and internet usage so, analyze, cross tabs, familiarity, and internet usage so, familiarity will go where column zero go to statistics chi-square continue in cells percentage is column-wise continue and click ok, and what are we getting now? So, what you are getting?

0.01 which is less than 0.05 which means there is a significant association. You reject the null hypothesis. Right? And accept the alternate hypothesis. 0.01 you see continuity correction.

0.01 is is less than 0.05. Right? So, less than equal to 0.05 which means the probability is only 1% which means our initial assumption of there is no association null hypothesis itself is wrong we reject the null hypothesis because the p-value here you look at it is continuity correction it is 0.010 less than equal to 0.05 if it is less than 0.05, which means

the probability is so small for obtaining such a sample, which means if in the initial if in the original population there is no association and then I obtained this sample it is only 1% chance this is what it is saying that it is 1% chance that I obtained such a sample where this difference is so huge.

And I cannot be so unlucky that I get this 1%. So I reject the initial assumption H0. That there is an association. So P value here is less than 0.05.

Continuity correction are you getting? 0.01. 0.01 is less than 0.05 or no? Are you sure? 0.01 is less than 0.05, right? So where is the confusion?

There is a video I wanted to show. Good afternoon student. This is the second video of the series wherein I am discussing the chi-square test. Now after you have done the data cleaning and the basic descriptive then plotting the graphs, this test comes if there are categorical variables and you want to see whether there is a relationship between categorical variables. Categorical variables are nominal in nature if you remember that.

So here we want to understand whether gender has any relationship with feeling stress in times of COVID. So gender is measured as male female and feeling stress as yes or no. And these are sort of 20 men and women, 20 male and female. And to do the chi-square analysis, we put the variable and the categories on X and Y axis. So we can do categorical with two variables

male and later male female yes and no, so how many males if you count? how many males have feeling stress in time of covid? 16 yes and 4 no and females 18 females 18 no, and if you count females are 18 are saying no and 2 are saying yes so 90% and 10% So, you see the percentages will give you some idea whether there is a relationship between gender and the feeling stress in times of COVID or there is a difference between the male and female in terms of feeling stress in times of COVID. So, 80% are feeling stress and 20% are not. 10% are not feeling stressed for females and 90% are not.

Now, by looking at percentages, you see that there is a difference between males and females. Now, is there a way to quantify? Yes, there is a cash flow test. Now, what cash flow test is? Cash flow test involves observed, this is the formula, summation of observed minus expected whole square divided by expected.

Now, what is the observed value? Expected is assuming that there is no difference between, there is no relationship or there is no difference. What will be the number in

these cells? So expected, assuming that there is no difference, the number will be 10, 10, 10, 10. So expected for this cell, observe this extreme, expected is 10.

You can calculate that and do it for all four cells and calculate the number, that will be chi-square number. And in chi-square number, whatever you get, you have to see the chi-square table. At 95% level of significance, if that value is beyond a critical value, then you say that there is a difference between feeling of stress between males and females. If that particular chi-score value is within the critical value, less than the critical value, then you say there is no difference,

by computing observed minus expected for everything 60. So for first one it would be 60 minus 10 divided by 10 plus then this plus this this you will get a chi-square value suppose you get 25 chi-square value and chi-square is and critical value in the chi-square table suppose 3 and 25 is well beyond 3, then it means this value is significant there is a difference between males and females, feeling of stress in times of COVID. So this is the Casper test. Thank you. So the next one familiarity internet usage you have done.

Now before we move on to the next sheet, wherein we are going to begin with the hypothesis testing for differences. So hypothesis testing as I said there are two types, one is, test of association, and test of differences, test of differences where we use t-test independent sample t-test, and so on that we used in test of differences I want to draw your attention to sampling distribution, you know what is the sampling distribution? What is the sampling distribution? So let's say I want to know the average height of B.Tech students at IIT. How many B.Tech students are there?

Let's say 4000. I want to know the average height of BTEC students, I cannot go to 4000 and not possible everybody will not be willing to right? I'll be kicked so many times so what I do is I randomly draw sample 30 of size 30 and let's say I calculate the average height of that group is 5 feet 5 feet and let's say 4 inches. One sample I draw this is the mean height. Second sample I draw 5.5 inches.

Third sample I draw 5.4 inches. Fourth sample 5.6 inches. Fifth sample 5.8 inches. If I keep drawing this sample infinite number of times I will end up with a curve like this. Which means for each sample of size 30, if I calculate the mean, I will get most of the samples, the mean would be around the population mean, which is called mu.

Most of the samples would have, most of the samples would be here, their mean would be, so this is the frequency, most of the samples would have this mean, which is the

population mean, and then so let's say this is 5.6 and this is some samples I will get greater than, some sample I will get less than, and this is 95 percent of the area let's say this is called a sampling distribution entirely which means when I when I draw a sample size of 30, and take the mean of sample their height and I keep doing that 30 sample size from population of 4000 I keep drawing and draw this frequency curve I will get a frequency curve like this wherein in the middle the highest frequency will be the highest the mean with the highest frequency will be the population mean, whereas on sideways the frequency will be lower and lower and like this this becomes the sampling distribution I am plotting the mean of the sample drawn, and I will get a curve like this with most of the samples of the size 30 with the mean same as the population mean the actual mean of the population this is called sampling distribution, and why I am telling you this because this we use in hypothesis testing of differences quite a bit so in the sampling distribution the mean of the in the sampling distribution we draw the means for a particular characteristics

for a particular characteristic in a population and we try to see that the sample size is 30 or greater than 30 when the sample size is 30 or greater than 30 the sampling distribution will be near normal whereas, when the sample size is less than 30 the sampling distribution will be not normal and many of our statistical test have the assumption of the sampling distribution as the normal so once we understand the sampling distribution now we can go a little bit further to see how it is used now let's say that I tell in this class let's say, that uh since I am your instructor I can tell anything, I can brag about anything, and I assume that there is some level of trust and you might believe that, let's say, I tell you that at my house there is a backyard where the coconut trees are there and the average height of trees average height is 100 meters let's say, I tell 100 meters and you say okay, this instructor has been teaching us for long. Till now he has not lied.

Let us some of you would say 100 meters is too much. It's okay. He is saying let it be. Might be true. There is okay.

Now one among you is very smart. He says that I cannot let him do this. I cannot let him lie. So he wants to do hypothesis testing. Now, in hypothesis testing, as I said, he assumes that H naught, let us say, because I am going to do hypothesis testing, but since he taught me, I assume that what he says, let me assume that null hypothesis is correct.

Let me assume that null hypothesis is that the population mean average height of coconut trees in his backyard is 100 meters. So we do this. Innocent until proven guilty. Trust until proven that he is lying. And H alternate is mu is not equal to 100 meter.

So I assume that the average height is 100 meter. Who is that person? Then that person went to my home to test this hypothesis. Who is that person? So he went to do the hypothesis testing.

So his assumption, what is his assumption? His assumption is what instructor is saying is true. Assume that what he is saying is true. So which means if we draw, if we draw sampling distribution, new will be, assume that this is true. Which means assuming that instructor is true, what is the probability that,

so now, what he does is, once he goes to my home, he randomly selects 35 trees and he measures. He does a training before coming that how to measure the height of the trees and then he measures for a sample of 35. He measures it and he gets the average height as 9 meters. So 9 meter is there. So he says that instructor was saying that it is 100 meter average height.

I assume that this is the population mean. Assuming that he is right. Assuming that the null hypothesis is correct. What is the probability that in a population where the mean height is 100 meters, I am able to draw a sample with such a less mean height. What is the probability?

If the mean height is 100 meter, how unlucky I am to how unlikely it is for me to select a sample with the such a mean height 9 meter, so if I see it is 1% which means there is only 1% chance that you select such a sample it is so unlikely and you cannot be so unlucky that you get that particular sample right? so what you do is, if you assume that null hypothesis is correct, which is population mean is 100 meters, null hypothesis is correct, it is very unlikely that any sample would have, you know, mean height of 9 meters. So, if it is very unlikely, so what is your assumption? That population mean is 100 meters, which is wrong. So you reject the null hypothesis, and this is what we call this is 2.5 percent and this is 2.5 percent, if the value falls here or here which means there is only very 2.5 percent chance of something happening here, 2.5, 2.5, if it is, so unlikely, given that we assume null hypothesis correct

in those cases we reject the null hypothesis, so what we do is, we start with the assumption that okay instructor has said it is 100 meter, assume that it is correct, now if it

is correct let me draw a sample and its mean, how and how likely is that I am to get such a mean 9 meter, it is very unlikely, which means the initial assumption would have been wrong that is how we do hypothesis testing. Is it something clear now. Right? So this is the sampling distribution I am using to do the this is one sample t test we will be doing in the next class. So yeah, so we will look at here this this will be the way we will be doing one sample t test, where we will be seeing, whether the mean, whether the population mean is equal to some value, which assume let us say yeah. Yeah so so, what what I am saying is, in the next class we will use this to do a t test, let's say one sample t test wherein we'll say in a population let's say we want to see whether the mean is equal to that's for familiarity with the internet, we want to see whether the mean is equal to 6 or not so we will assume that. Let's say the mean is equal to 6, mu is equal to 6, what is the probability that from the sample the mean is let's say 2.5, if it is less than equal to 0.05 we reject the null hypothesis, if it is so we can use this particular sampling distribution this concept to do one sample t test, then we will use it to do independent sample t test, but sampling distribution and p value understanding is very important to

get, I hope to some extent you got it. Right? let's end the class here and we will meet in next class thank you.