Marketing Research and Analysis-II (Application Oriented)
Prof. Jogendra Kumar Nayak
Department of Management Studies
Indian Institute of Technology – Roorkee

Lecture – 30
Non-Parametric Test - VI

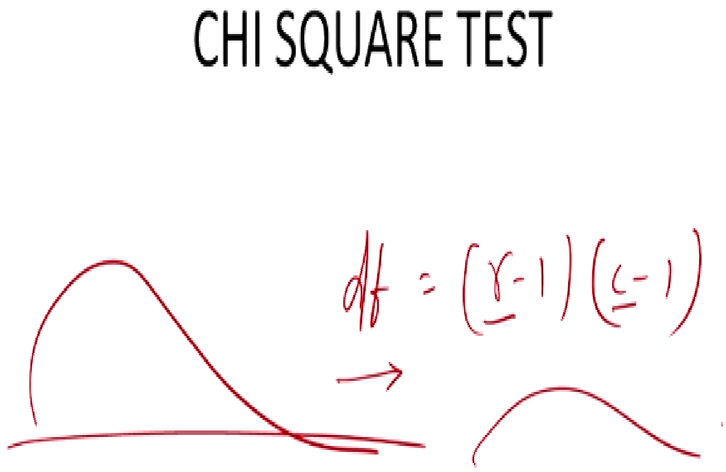
Welcome friends to the course Marketing Research and Analysis. I hope you are understanding the concept of nonparametric tests and its utility in research. Many a times people they try to ignore the nonparametric ones thinking they are less powerful, yeah that is not wrong, the nonparametric tests comparison to the parametric ones they are less powerful but then when the data does not have the requirements of a normal distribution and does not have the requirements of a parametric test, a parametric test cannot be done.

So you would have no option but you have to continue with the nonparametric tests and in that condition the nonparametric test is the most apt tests. So if a data is not normal, you can easily do some nonparametric test and can show it in your research work, in your publications, and may be in your thesis okay. So some of the nonparametric tests we have already covered are like the sign test, the Wilcoxon signed-rank test, the Mann-Whitney test, and the Spearman correlation test which is a test of association and talks about the strength of relationship between 2 variables which are non-normal in nature.

In the last lecture, we had stated with one of the most powerful tests called the chi-square test. The chi-square test as I said is a test of association between 2 variables which are nominal in nature and it helps to understand whether there is any relationship or any association between the 2 variables. So let us say in your questionnaire, many a times I have seen students collecting data like for example what is your religion and what is your let us say income or let us say what is your occupation and what is your income.

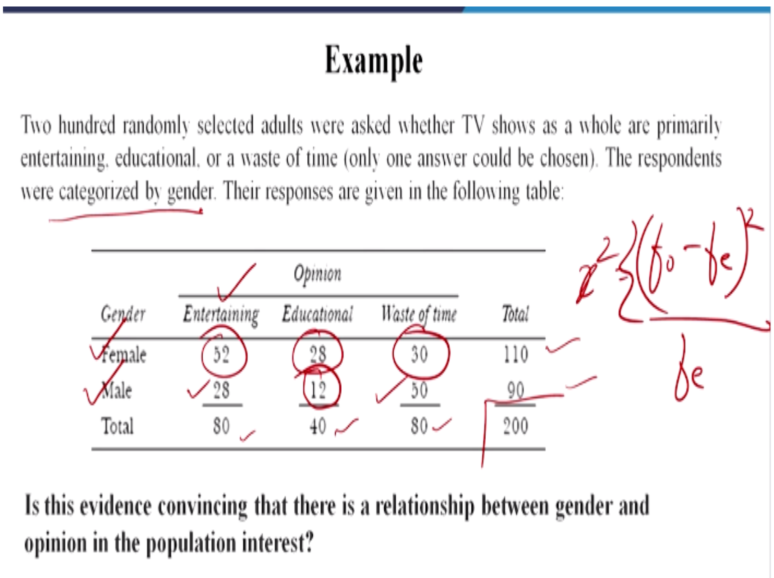
An income can be taken only a ordinal variable or so, or let us say what is your religion and what is your let us say which place you come from and your level of hygiene, now both are data which are nominal in nature, your country of origin is also a nominal variable and your level of hygiene is also a nominal variable. So in such a condition when you have such data, we should not be leaving the data and we can do some tests and check whether there is any association between such kind of variables. So in such conditions, the chi square test comes into play.

(Refer Slide Time: 03:18)



So we had discussed about the chi-square distribution which is a right skewed distribution and I said as the number of degrees of freedom which is calculated by the figure r -1 x c-1, so what is r, r is the row, c is the column. When you multiply these two factors it gives you the degree of freedom. So as the degree of freedom goes on increasing, the chi-square distribution which actually is something like this tends to become more of a normal in nature, so this is what we had said.

(Refer Slide Time: 03:39)



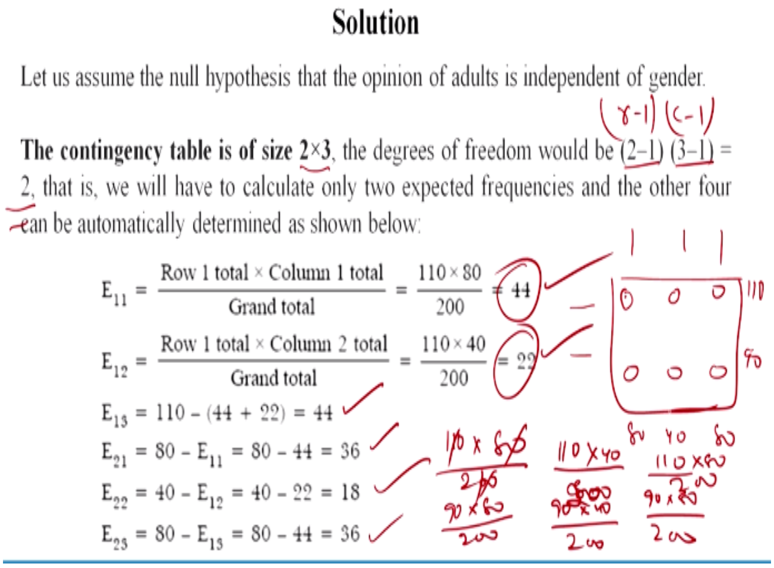
So today what we will do is we will start with an example of the chi-square test. So I have already explained the chi-square test and how you should be going about. So if you just remember the chi-square formula. So the formula was observed frequency minus expected frequency square, the summation of this, divided by the expected frequency. So this is what was the formula. So in this case let us take this problem, 200 randomly selected adults were asked whether TV shows as a whole are primarily entertaining, educational, or a waste of time.

So many a times in your homes also this discussion might be going on. Parents must be saying the TV is a waste because the children are wasting the time, somebody must be saying no no there so many educational programs so they are learning, or sometimes somebody says it is only just for entertainment. The respondents were categorized by gender, so female and male okay. The responses are given the in the following. Some people where 52 females said it is entertaining in nature, 28 males said it is entertaining in nature.

When it came to educations the opinion, 28 females said it is educational in nature, only 12 males said it is educational in nature. When it came to whether it is a waste of time or not 30 females felt it is a waste of time, TV is a waste of time, but on the same condition same case 50 males thought it is a waste of time. So these are all the observed frequencies and the total if you count 52+28+30 how much 52, 60, 70, 80+ thirty 110, 28+12+50 that comes to 20, 30, 40, 50, 90. Similarly this side 52 70 80, 40, and 80.

So these are all known to us and the overall total if you see this 80+40+80, 200; 110+90, 200. So the question here is, is this evidence convincing that there is a relationship between gender and opinion in the population interest. So what do you think, from watching this data you might be able to predict something but that cannot be justified. So we have to statistically check and we know that this data is just a nominal data, so a normal distribution does not hold true. So parametric tests are not allowed. So how do we go further/

(Refer Slide Time: 06:36)



Let us assume the null hypothesis that the opinion of adults is independent of gender, that means gender has no impact, male and female, so what male says and what female says there is no difference, it is more or less the same. Now we build contingency table. So what is this contingency table we said. There 2 genders, male female, and 3 opinions, so a table is formed which is called a contingency table of 2 x 3. The degrees of freedom is how much, now r-1 x c-1, so 2-1 x 3-1 = 2.

So we will have to calculate only 2 expected frequencies and the other 4 can be automatically determined. let us see how. So the expected frequency for the first one first cell, let us say this this one, is row total which was how much, let us bright it here,110, 90 that was 80, 40, 80 okay so this case, so how much it is 110 x 80/200. Now this case how much, 110 x 40/200, this case 110 x 80/200, similarly this case 90 x 80/200, this one 90 x 40/200, third one 90 x 80/200.

Now if you see so the expected value for each cell is now given, 44, 22, so 110 x 80/200 let us see so 11 x 8 are 88/2 44 right, so similarly 22, now e E13 is 44, this is 36, so you can just simply calculate, 18 and 30, so every cell can be calculated and now you have got the expected value.

(Refer Slide Time: 08:54)

The contingency table of expected frequencies is as follows:

	Opinion			
Gender	*Entertaining*	*Educational*	*Waste of time*	*Total*
Male	44	22	44	110
Female	36	18	36	90
Total	80	40	80	200

So the now the expected value frequency looks like this, male who feels it is entertaining it is 44 and female 36. Educational, 22 males feel it is educational, 18 females feel it is educational as per the expected frequency. Waste of time, 44 and 36. So this is there. Now we will calculate the chi-square value.

(Refer Slide Time: 09:21)

Arranging the observed and expected frequencies in the following table to calculate the value of $\chi2$ -test statistic:

Observed (*O*)	*Expected* (*E*)	(*O* − *E*)	$(O-E)^2$	$(O-E)^2/E$
52	44	8	64	1.454
28	22	6	36	1.636
30	44	−14	196	4.455
28	36	−8	64	1.777
12	18	−6	36	2.000
50	36	14	196	5.444
				16.766

Now this is the observed value, we have this, which is expected we have calculated. So observed minus expected is equal to 52-44, so how much 8; 28-22, 6; 30-44, 14, so this should be minus; this is 28-38, -8; this 12-18, -6; 50-36 is 14. So now we cleared the square of it. So 8 square 64; 6, 36; 14, 196; 8, 64; 36; 196. Now the final result, this is what of interest was is a chi-square. So chi-square is this, observed minus expected square divided by the expected, so that is 44 in this case right, divided by expected is 22 in this case.

So when you do, you get these values correct, so each one of them. So total chi-square value is 16.766 okay.

(Refer Slide Time: 10:28)

The critical (or table) value of $\chi2$ = 5.99 at α = 0.05 and df = 2.

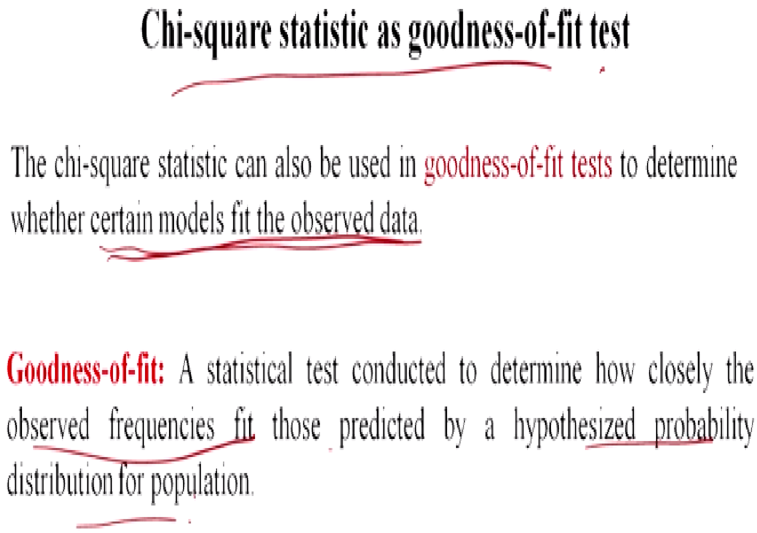
Since the calculated value of $\chi2$ = 16.766 is more than its critical value, the null hypothesis is rejected.

Hence, we conclude that the opinion of adults is not independent of gender.

So the critical value for chi-square is equal to 5.99, when the alpha level is 5% or 0.05, for the chi-square distribution, the table value is at a degree of freedom 2 = 5.99, that you can check from any chi-square distribution table taking the degree of freedom as 2 and level of significance as 5, you will find it is 5.99. So since the calculated value of chi-square is, so it is something, so your calculated value is here somewhere let us say is 16.766 and what is your chi-square tablet value, now only five 5.99.

So it is more, it is going way beyond. So the null hypothesis is going beyond the acceptance region and it is into the rejection zone, so this is the rejection zone from this side, so the null hypothesis is rejected. What does it mean? It means we conclude that the opinion of adults is not independent of gender, that means we can say easily that the opinion of male and the opinion of female are not same, but they are different. So this is the first utility of the chi-square test.

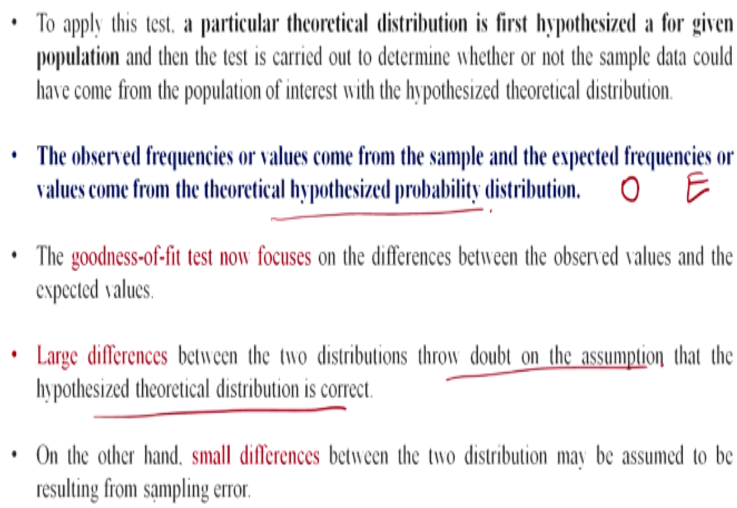
(Refer Slide Time: 11:52)



Now chi-square test is also used as a goodness-of-fit test. So what is this goodness-of-fit let us see. The goodness-of-fit is a very important statistical utility. The chi-square statistics can be used to determine whether certain models fit the observed data. In fact when you talk about structural equation, modeling, and higher statistical tests, you will realize that the basic algorithm is based on again the chi-square where the expected model and the observed model are compared and lesser the difference the more the stronger the model and higher the difference the weaker the model.

So this is used in many such statistical tests. So what it says? It helps in understanding, it determines whether the models fits the observed data. So whatever data you have observed you have got in hand, whether your model is fitting to the data or not. So what is this goodness-of-fit again. A statistical test conducted to determine how closely the observed frequencies fit those predicted by a hypothesized probability distribution or an expected one, let us say.

(Refer Slide Time: 13:01)



To apply this test, a particular theoretical distribution is first hypothesized for a given population and the test is carried out to determine whether or not the sample data could have come from the population of interest. So you first hypothesize and then you try to see whether it has come from the same population or it has not come. The observed frequencies come from the sample and the expected frequency value comes from the theoretical hypothesized probability, so that I have said is a comparison between the observed versus the expected.

The goodness-of-fit now focuses on the differences between the observed values and the expected values. Large differences between the 2 distributions throw doubt on the assumption that the hypothesized theoretical distribution is correct, that means that it has come from the same population this hypothesis is doubtful in nature. On the other hand small differences or else they are close equal, then the 2 distributions may be assumed to be resulting from sampling error.

(Refer Slide Time: 14:03)

Procedure

The general steps to conduct a goodness-of-fit test for any hypothesized population distribution are summarized as follows:

Step 1: **State the null and alternative hypotheses**

H_0: No difference between the observed and expected sets of frequencies.

H_1: There is a difference

Step 2: Select a random sample and record the observed frequencies (O values) for each category.

Step 3: Calculate expected frequencies (E values) in each category by multiplying the category probability by the sample size.

So this is what it talks about. So let us check the goodness-of-fit test. The general steps to conduct a goodness-of-fit test is as given. First state the null and alternative hypothesis. So what is the null and alternative hypothesis. Null is no difference exists between the observed and the expected set of frequencies, so there is no difference. Alternate says there is a difference between the observed and the expected. Now select a random sample and record the observed frequencies for each category, calculate the expected frequency in each category by multiplying the probability by the sample size. We will do that, we will see a problem.

(Refer Slide Time: 14:43)

Step 4: Compute the value of test statistic

$$\chi^2 = \Sigma\frac{(O-E)^2}{E}$$

Step 5: Using a level of significance α and $df = n - 1$ provided that the number of expected frequencies are 5 or more for all categories, find the critical (table) value of x^2

Step 6: Compare the calculated and table value of $\chi 2$, and use the following decision rule:

- Accept H_0 if x_{cal}^2 is less than its critical value $x^2_{\alpha, n-1}$
- Otherwise reject H_0

Now you have calculated the same thing. Now using a level of significance and degree of freedom provided that the expected frequencies are 5 or more at least, so each cell has to have at least 5 values, the size has to be 5 or more. Now compare the critical calculated and the table value and then we do the same thing as earlier.

(Refer Slide Time: 15:07)

Example

A Personnel Manager is interested in trying to determine whether absenteeism is greater on one day of the week than on another. His records for the past year show the following sample distribution:

Day of the week :	Monday	Tuesday	Wednesday	Thursday	Friday
No. of absentees :	66	56	54	48	75

Test whether the absence is uniformly distributed over the week.

Let us take this case. A personnel manager is interested in trying to determine whether absenteeism is greater on one day of the week than on another, that means is the attendance the same throughout the week or is it that in one particular day, may be it is a Monday, where people are lazy enough to not to come on time or is a Friday or something. His records for the past year show the following sample distribution.

Monday number of absentees 66, Tuesday 56 people absent, so as I said may be after the Sunday people have not yet come out of their the laziness is still in there. So Wednesday 54, Thursday forty 48, now Friday is interesting again 75, so some people have may the Saturday is an off so they have taken extended holiday and taken a Friday as a holiday also. Test whether the absence is uniformly distributed over the week or not.

(Refer Slide Time: 16:06)

Solution

Let us assume the null hypothesis that the absence is uniformly distributed over the week.

H_0 = Uniformly distributed.

The number of absentees during a week are 300 and if absenteeism is equally probable on all days, then we should expect 300/5 = 60 absentees on each day of the week.

Now arranging the data as follows:

Let us assume the null hypothesis is that the absence is uniformly distributed. So obviously H0 is uniformly distributed. The number of absentees during a week are 300 and if absenteeism is equal, then it should be how much, 300/5 = 60 on each day. So just let us see how this, you add it up, 60+50, 110, 160, 200, 270, 275, 284, 287, 293, 299, may be I think make it one of them it is not exactly 300, it was approximate 299, so we are going to assume it is 300.

(Refer Slide Time: 16:55)

Category	O	E	$O-E$	$(O-E)^2$	$(O-E)^2/E$
Monday	66	60	6	36	0.60
Tuesday	57	60	−3	9	0.15
Wednesday	54	60	−6	36	0.60
Thursday	48	60	−12	144	2.40
Friday	75	60	−15	225	3.75
	300	300			7.50

The critical value of $\chi 2$ = 9.49 at $\alpha = 0.05$ and $df = 5 - 1 = 4$ ($n-1$)

Since calculated value x_{cal} = 7.50 is less than its critical value, the null hypothesis is accepted.

Now arranging the data you see how it is coming, now 66, 57, it is written over there, this was I think there was a mistake here, so this is 57 actually, so we can see here 57, so that comes 300, this is also 300, so equally divided now. So this when you calculate the O-E is 6, -3, -6, -12, -15. So this O-E square we have calculated and divided by the expected, this is the chi-square value. So we have got 7.5 chi-square value. So the critical value of chi-square at

5% level of significance with a degree of freedom at 5-1 = 4, this is nothing but the n-1, it is 4, is how much in a 9.49.

So since calculated value is 7.5 and the critical value is 9.49, the null hypothesis is within the acceptance region, so it is to be accepted that there is no difference or there is uniformity. So what is the null hypothesis saying, the absence is uniformly distributed, now yes the absence is uniformly distributed, that is what the null hypothesis says and it has been accepted. So this is what we talked about a chi-square test. What I will do is I will try to show you how to conduct a chi square test on the SPSS.

(Video Starts: 18:22) Now let us look at this data. Now the same data which I was talking about. There are 2 genders, for example let us go to the variable view. Now gender 0 is female, 1 is male and when you come to opinion 1 is entertaining, 2 is educational, 3 is waste of time. Now what are we saying. To do a chi square test, we need to cross check whether there is any association between opinion and the gender, that means do females speak differently and males speak differently or they are similar.

So do that, just go to analyze, go to descriptive statistics and you find here crosstabs. So when you take crosstabs, you take these 2 into the 2 column and row, it does not matter which one you take where, it actually does not matter. So now let us go to the statistics. So I need the chi-square. So you can see there are several things for example the Phi coefficient, contingency coefficient, but I am not requiring all these things, so I will just go to with the continue. So now let us see what is the test value which has come.

Now when you have taken the case processing summary says N is 60, so there are no missing values, 100%, and gender opinion cross tabulation it says female who said entertaining is 7, male 14, so total 21; educational 15 and 8, 23; waste of time 5 and 11, 16. So total female we have 27, total males are 33. Who said the total of entertaining opinion is 21, educational 23, and waste of time is 16. So is there any association, can we say? Now let us check the Pearson chi-square.

Now when you look at the Pearson chi-square with a degree of freedom of 2 and you have a 0.46, is a two-tailed test, so when the p-value is 0.046, then it means that the null hypothesis is lesser than the 5% level of significance, so the null hypothesis is rejected. Now if it is rejected **(Video Ends: 20:48)**, what do we mean in this case? In this case, we mean that if it is rejected that means there is a difference between the opinion regarding entertainment, educational, and waste of time between male and female.

So this is what we have learned through this output and this is how you should write that at particular degree of freedom and level of significance, the table value is this much and the calculated value is this much, since the table value is lesser than the calculated value so the calculated value is more and the p-value is more, it is to be rejected.

(Refer Slide Time: 21:24)

Now we come to another test called the McNemar test. Now McNemar test is a very interesting test, it is very similar to the chi-square test because it follows a chi-square distribution, but the only thing is this McNemar test is a two 2 x 2 structure.

(Refer Slide Time: 21:42)

What is the McNemar Test?

- The McNemar test is a non-parametric test for paired nominal data.
- It's used when you are interested in finding a change in proportion for the paired data.
- For example, you could use this test to analyze retrospective case-control studies, where each treatment is paired with a control.
- It could also be used to analyze an experiment where two treatments are given to matched pairs.
- This test is sometimes referred to as **McNemar's Chi-Square test** because the test statistic has a chi-square distribution.

It is a nonparametric test for paired nominal data, used when you are interested in finding a change in proportion for the paired data. For example you could use this test to analyze retrospective case control studies where each treatment is paired with a control, that means before and after, let us say case of before and after. What happened, what was our health, how was your health, your weight before you took some protein diet and what was your health after you took some protein diet.

It could also be used to analyze an experiment where 2 treatments are given to the matched pairs. This test is sometimes referred to as McNemar chi-square test because it has a chi-square distribution okay.

(Refer Slide Time: 22:30)

Assumptions for the McNemar Test

The three main assumptions for the test are:

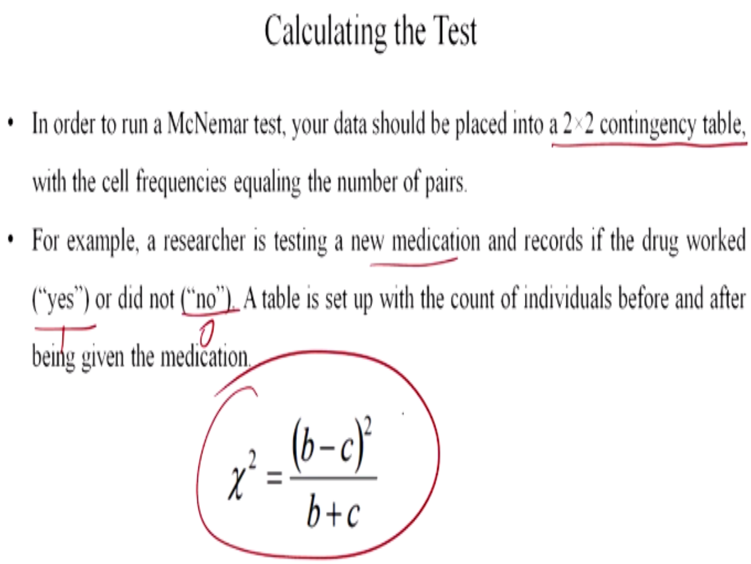
- You have one categorical dependent variable with two categories (i.e., a dichotomous variable) and one categorical independent variable with two related groups. Examples of dichotomous variables include perceived safety of helmet(two groups: "safe" and "unsafe"), exam performance (two groups: "pass" and "fail").
- The two groups in the dependent variable must be mutually exclusive. In other words, participants cannot appear in more than one group.
- Your sample must be a random sample.

Note: *If your data does not meet these three assumptions, considering running another test for your data like a regular chi-square test.*

So what are the assumptions? The 3 assumptions are you have one categorical dependent variable with 2 categories, that is it is a dichotomous variable, and 1 categorical independent variable with 2 related groups. So as I said, it is 2 x 2 matrix basically. Now this is an example of dichotomous variables include perceived safety of helmets. So 2 groups are there one who feels safe and the other one is unsafe, so safety of helmet. Now exam performance for example 2 groups, pass or fail, so these are the 2 dichotomous variables.

The 2 groups in the dependent variable must be mutually exclusive, in other words participants cannot appear in more than one group, suppose if you are pass then you cannot be fail obviously logically. So your sample must be random sample. Note; if your data does not meet those 3 assumptions, considering running another test for your data like a regular chi-square test is preferred.

(Refer Slide Time: 23:29)



How do you run the McNemar test. Your data should be placed in to a 2 x 2 contingency table with the cell frequencies equaling the number of pairs. For example, a researcher is testing a new medication records if the drug worked, yes if it worked, no it did not work. A table is set up with the count of individuals before and after being given the medication. So when before the medicine what was the condition, after the medicine was given what is the condition. So if you see, it looks like a chi-square right, so what is the formula saying b-c square/b+c.

(Refer Slide Time: 24:08)

Drug 1

Drug 2		No	Yes	
	No	80 a	100 b	180
	Yes	10 c	110 d	120
		90	210	300

$$\frac{(b-c)^2}{b+c} = \frac{(100-10)^2}{100+10} = \frac{(90)^2}{110}$$

So these are the values of the cell okay. Now drug 1, drug 2. Now what was the effect, no yes, no, yes, so there are 4 conditions. So we are not interested with this no no and yes yes case.

(Refer Slide Time: 24:27)

Cells b and c are used to calculate the test statistic; these cells are called "discordant." The other pairs a and d do not tell us whether the treatment is helpful or not.

- The McNemar test formula is:

$$\chi^2 = \frac{(b-c)^2}{b+c}$$

For the set of data above, we have:

$= (100\text{-}10)^2 / (100 + 10)$

$= 90^2 / 100$

$= 73.63$

So if you see b and c, these two, what it said are used to calculate the test statistic and these are called as discordant, this is not required, this is not be used. This no, yes, yes, no so that is where the difference on thought process lies. The other pairs a and b do not tell us whether the treatment is helpful or not, obviously it is both saying the same thing no, no, yes, yes, so does not tell us any difference. So the McNemar test is what b-c, now go to this b-c square/b+c, so what is the value.

From the above said b is how much 100, b-c square/b+c that equals to 100 minus c is how much 10 square upon 100+10, so that is equal to 90 square/110, so much it has come, so

73.63. Now you can compare it with the value and see whether this is actually the hypothesis is coming true or not or the hypothesis is to be accepted or not. So there are several cases where McNemar test comes to use. It is as good like a chi-square test, but only problem is that it is only a 2 x 2 case and it cannot handle a larger contingent table like a 3 x 3 or 2 x 4 or 2 x 5 it cannot handle, that is the only difference.

Well, what we will do is in the next lecture, I will show you a few more tests how to cover them on the software, for example the McNemar test, I had forgotten to do the Wilcoxon test also signed-rank test and then I will follow the case test and then nonparametric test gets over there. Then we will start with the next from there, the experimental design. So well, thank you for today and we will meet in the next class. Thank you so much.