INDIAN INSTITUTE OF TECHNOLOGY ROORKEE

NPTEL

NPTEL ONLINE CERTIFICATION COURSE

Marketing Research

Lec -22
Cross Tabulation and Chi- Square Test

Dr. Jogendra Kumar Nayak
Department of Management Students
Indian Institute of Technology Roorkee

Welcome friends once again the section of marketing research and analysis in our pervious section we will discuss about one of the most important techniques of you know testing hypothesis in which we are covered the analysis of variance becomes is a highly unutilized in all kind of studies basically beat experimental or non experimental or partially experimental whatever it is.

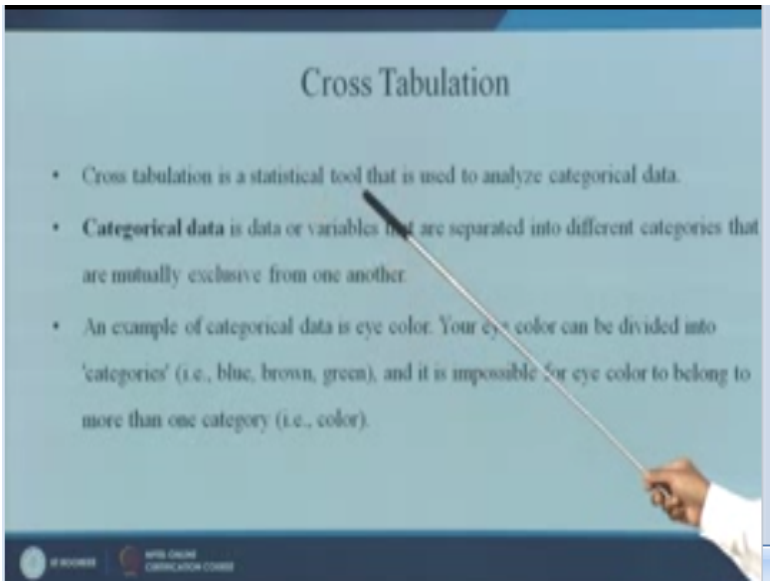
So it has a lot of unity and it helps you to study among variables which are both you know categorical as well as continuous right but what if suppose, you have a situation where there are data in which only let us categorical data are available nominal data or only available or ordinal data or only available right.

So if the data is only nominal in ordinal we generally have a tendency to I have see the students and others to you know say only percentage so 40% people this 50% people and they leave it there right the research so the question is there is way that we can use this type of re technique which can help us in analyzing data when it is in only a non metric nature right so let us in fact the technique now I am going we are going to discuss is basically one of the most utilized techniques in marketing research .

In fact if one the estimates such said that around 90% of the researchers in marketing research use this technique that we are going to discuss right now so what is this technique we are going to discuss so it is nothing but called the cross tabulations right so cross tabulation is nothing you in a tabular format you put in the data and see key how they what is there pattern okay right.

So and $\chi 2$ test being one of the cross tabulation is a part of the cross tabulation only it is a test it is a very powerful test which is used in all most of the studies where it helps you to make a statistical hypothesis testing right which was earlier if which if one does not know this then you would not be able to do hypothesis testing with nominal data or ordinal data right so this as a very large importance okay.

(Refer Slide Time: 02:56)



Let us slowly to the study so what you are saying cross tabulation is a statistical tool that used to analyze categorical data, here categorical data are variables is data or variables that are separated into different categories or which are different from each other mutually exclusive from one and another let us say one of the categorical data is eye color, so your eye color can be divide into let us say blue, brown, green, black okay which is normally let us say black and it is impossible for one eye color to belong to more than one category.

So that is why it is called categorical it can be staying in one category and very seldom where rare even if it is 2 category then it is in more in nominal another it becomes a category ordinal data may be it can fall into 2 groups right.

So in the earlier case as I said T test and ANOVA we where determining as relationship between the nominal variable and a continues variable right but what if we are interested in the 2 nominal variables right example class and unemployment right gender and drug use let us say that city

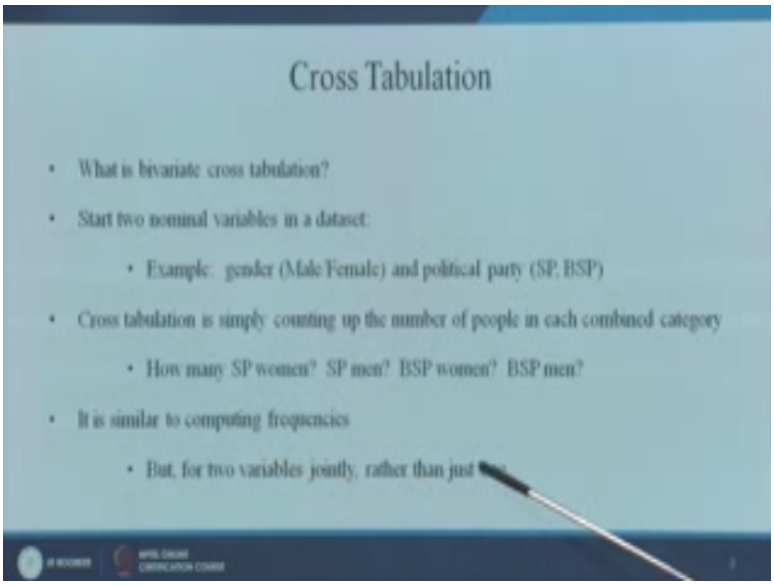
and likening of which city you come from and which player do you like in a particular foodball team or a cricket team right.

Or which city various which team okay so or let us which gender what gender likes what kind of TV programs so there are studies which might not be always the data might not be continues in nature right so continues is one thing but what if is there a data is both in nominal right and there was saying so let us say as I said okay let us say the data is you want to find out a relationship between let us say gender which is categorical let us only male and female and let us say the kind of TV program.

(Refer Slide Time: 04:50)

Let us say TV program could so drams right or spots okay or reality shows okay so whatever now is a relationship suppose a manager can know about this thing and automatically it can be of great use to him or her by creating a better policies better plans accordingly for the market okay so cross tabulation is use as nominal or ordinal variable.

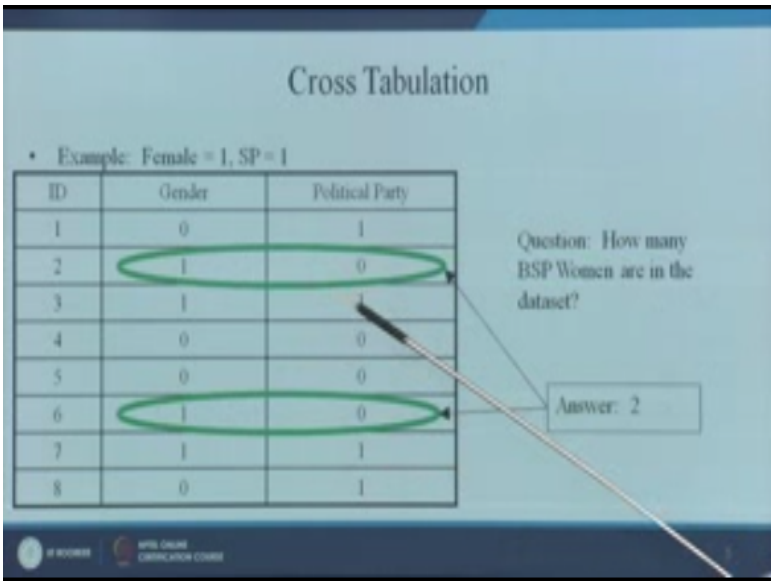
(Refer Slide Time: 05:23)



Right it is a toll to descriptively examine variables right tools it is uses tools to identify and there is a relationship between two variables, so basically it is sometimes you also called a test of a independent right we will see each one so what is bivariate cross tabulation that means by variant two variables starts with two nominal variables in a data set example gender and political parts so this is example I have brought so let say gender male female and political party could be let say we have two political parts is SP BSP right.

So cross tabulation is simply counting up the number of people in each combined category how many SP women how many SP men BSP women BSP man right, it is similar to computing frequencies but let see but earlier when we are doing a frequency generally about 90% of students do they only create a ϕ chart or an histogram show the frequency and that is all right but for two variables jointly rather than just one.

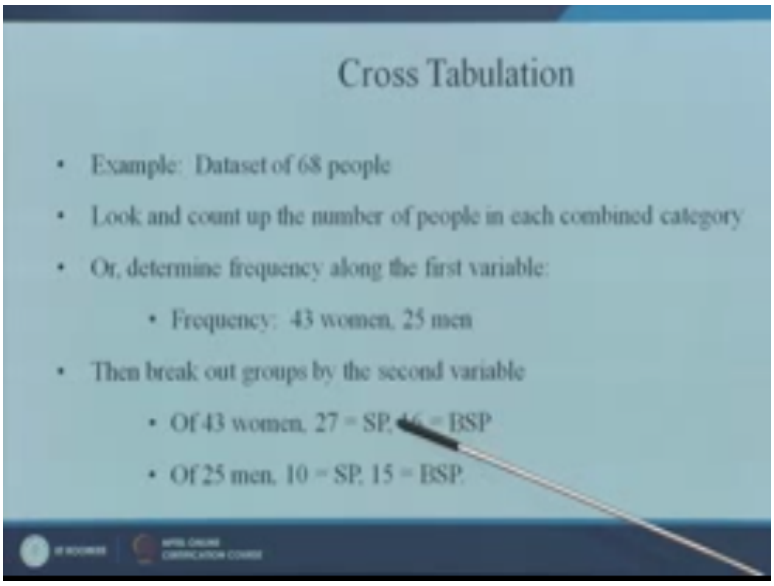
(Refer Slide Time: 06:25)



So let us see this let say female is equal to 1 and SP = 1 right the out of the two part is SP now this is a this is how the data looks like it looks like some like dummy coding right so 0 1 10 1 0 01 so this is how you have got the data for example now the question comes you how many BSP women are there in the party or in the data set now what is BSP now SP = 1 so BSP = 0 so 0 okay 0 0 0 so are this all 0s women now let us see now women was 1 so 0 1okay 1 this is 0 but this is 0.

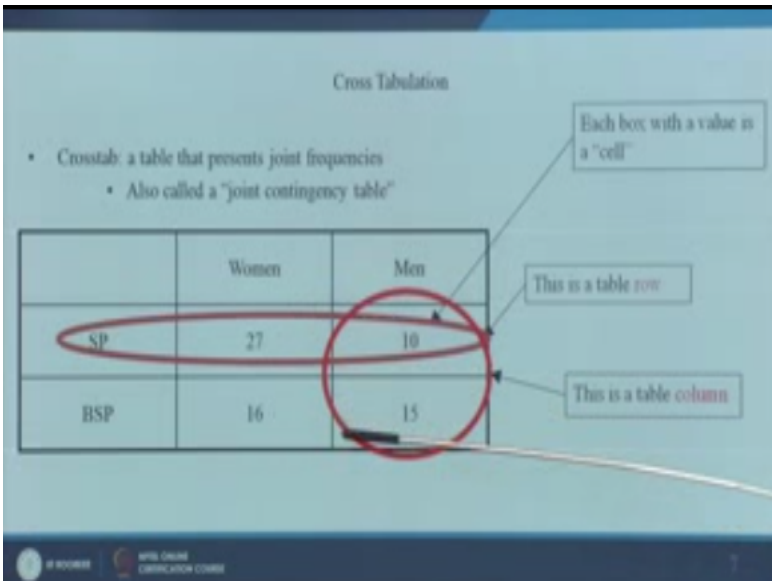
So he is a male right not possible 0 0 not possible 0 1 so there are two right, so these two answer is 2 okay.

(Refer Slide Time: 07:22)



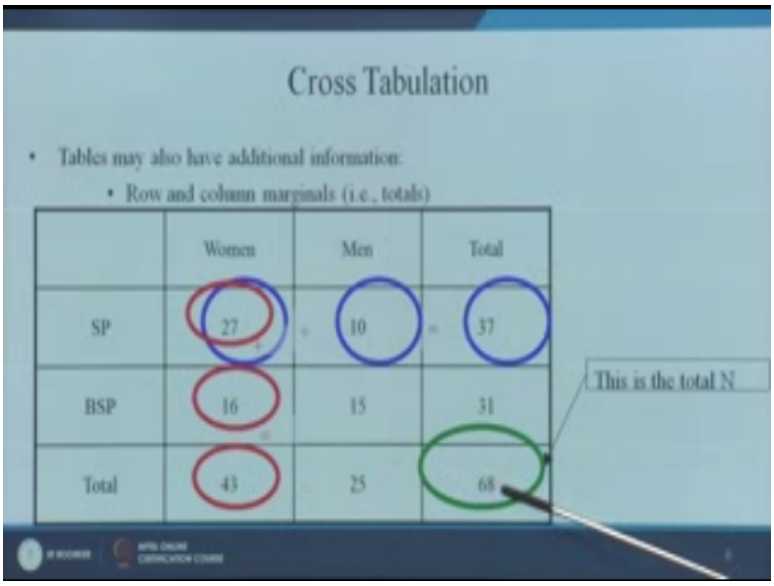
Now this is another this is let say this is a break up given so let say there are 43 women and 25 men of which 43 women are in towards 27 or in 2 SP 16 into BSP 25 men 10 is for SP and 15 is for BSP let say this is a break up okay.

(Refer Slide Time: 07:39)



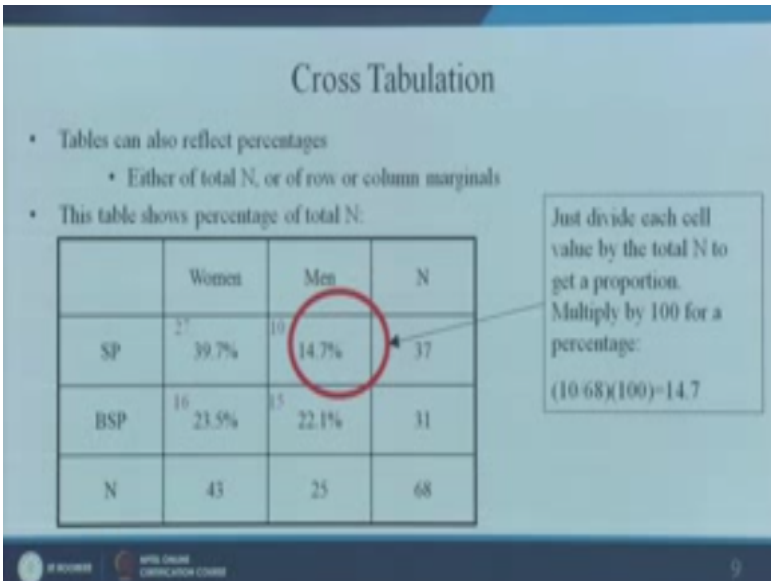
So how does a table joint table look like so it is also call a contingency table please remember so SP women SP men BSP women BSP men okay, so 27, 10, 16, 15 okay so this is a each box with the value is called a cell please remember right and what is the total now this is a table row right, this is a table and this is a column so this is the column so these are the table these are two rows and two columns okay.

(Refer Slide Time: 08:17)



Now look at this what it is saying so 27 + 10 so SP has got total 37 okay 27 women 10 men so total 37 BSP has got 16 women and 15 men so 31 so total SP is 37 total BSP is 31 total women is 33 total men is 25 so when the data is in such a format it is very, very helpful for the marketing researchers marketing researchers okay, so now let us this is a column and row and the column row total okay and this is the entire grand total, so the grand total is 37 + 31 is 68 or you move from this side the same 68 okay.

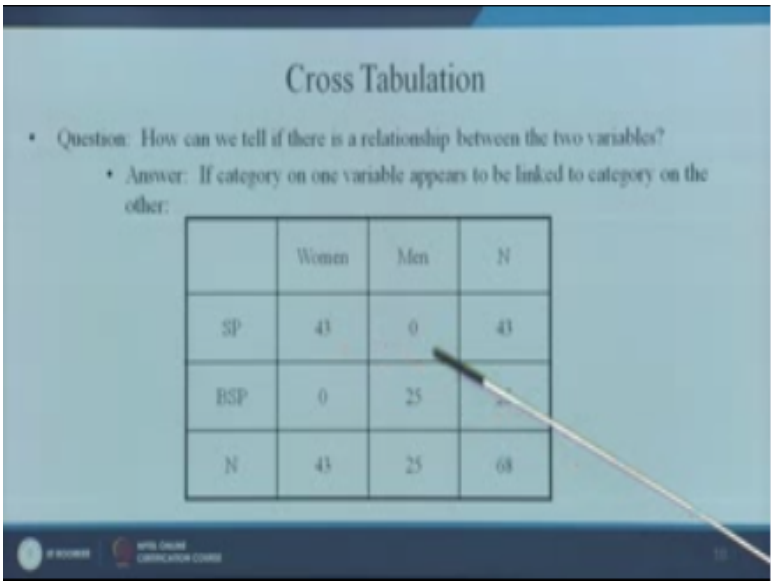
(Refer Slide Time: 09:02)



Now if you now find the percentage now let say this was 27 so 27 was what percentage of the total then 27 / 68, so 39.7% 14.7% so from this one can say okay what is the percentage of women who are what is a percentage of whether men or more in SP or women or more in SP let say and similarly for BSP now we can easily say that more women are there in SP and in case of BSP also it is women only right, it is women only but men or pretty higher in the case of BSP then SP right.

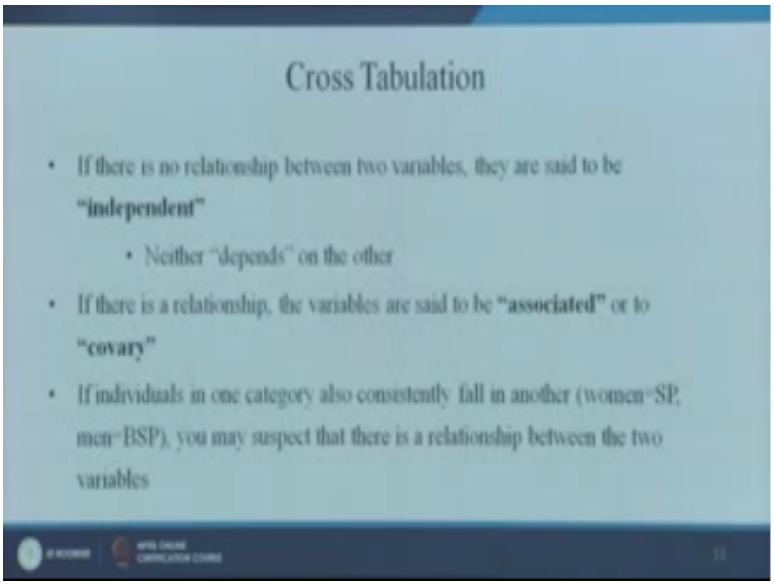
So although this is less but this difference among these two is not very high right but this difference is very high that means SP now has to try to get in more men than women right okay so this is we discussed right now how can we tell if there is a relationship between the two variables.

(Refer Slide Time: 10:00)



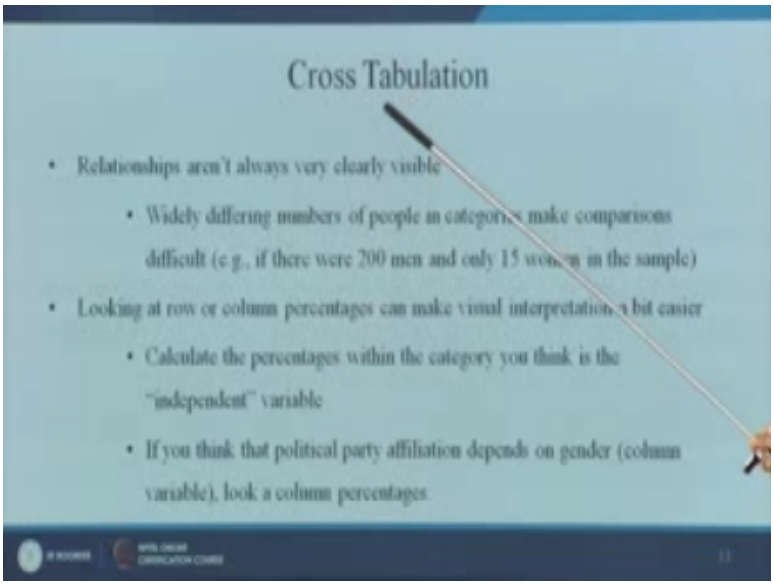
Now the question is till now we have discussed only from the frequency by an observation we said okay, this is what is the data now this from his percentage we can say okay what can be done right, but the question comes here can we tell that there is a relationship between the two variables, okay now let us see. If category on one variable appears to be linked to the category of other right, now let us check this. Now this is the case, now 43, 25 right, so totally 68 right.

(Refer Slide Time: 10:30)



Now if there is no relationship between the two variables they are said to be independent so many a times we say it is a test if independents right, so neither depends on the other that means if there is a relationship the variables are said to be associated or to covary, so χ^2 test works as a test of independence also, it works as a test of association also, right. If individuals in one category also consistently fall in another that means a person is women=SP and men=BSP, you may suspect that there is a relationship between the two variables that means what did he saying if individuals in one category also consistently fall in another that means they are visible in the both places. Then, there is a relationship between the two variables right.

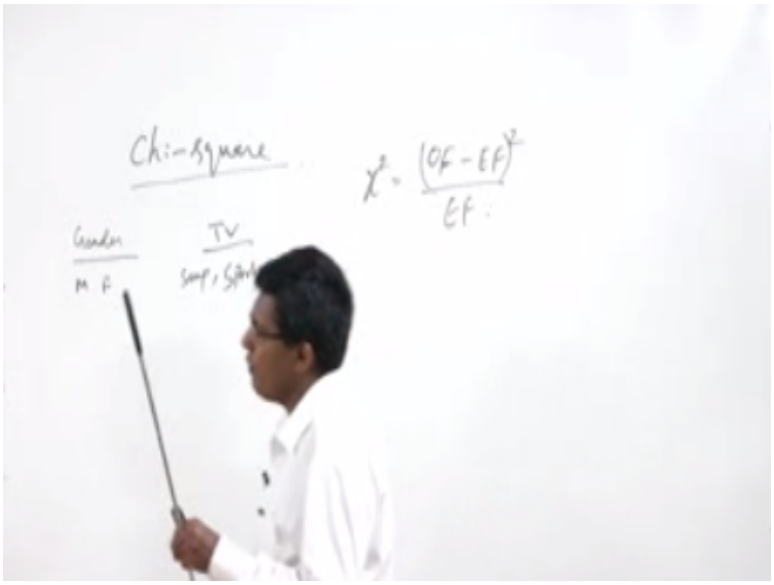
(Refer Slide Time: 11:19)



Now cross tabulation is says the relationships are always not very clearly visible right, so widely differing numbers of people in categories make comparisons very difficult example if there are 200 men and only 15 women the comparison would have been very tuff to do, right. Now you have done this through let us say we have gone it, gone through it so we have calculated the percentage and we will see given the test are independent or not, so and we thought from the value okay this is there is a, there are not dependent and there is a clear distinction so there independent of each other, okay.

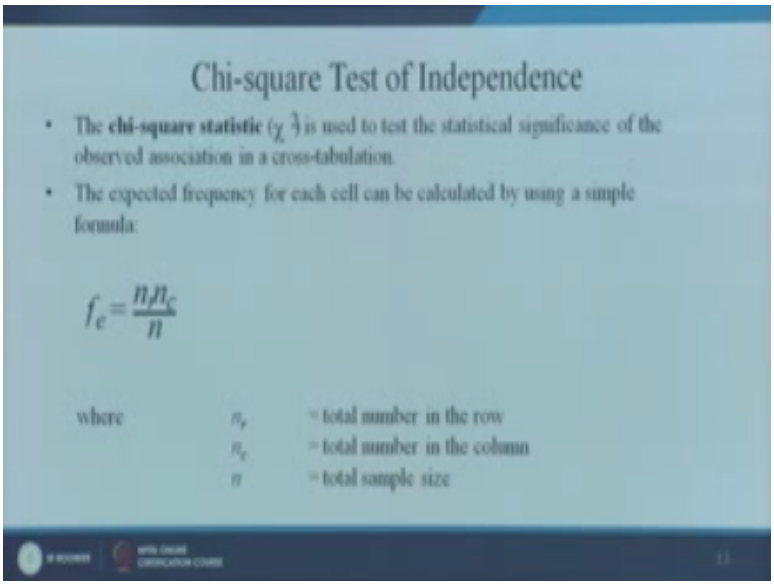
Now to test okay whether the relationships actually exists or not is there is a significant relationship or not just looking at a value will not help us, so we need to so have some hypothesis, some statistical test for it, so the statistical test which we use is call χ^2, χ^2 test is nothing but a test use for the cross tabulation or the data in the cross tabulation and it helps you in finding out whether that is in a significant relationship between the variables or not, okay.

(Refer Slide Time: 12:36)



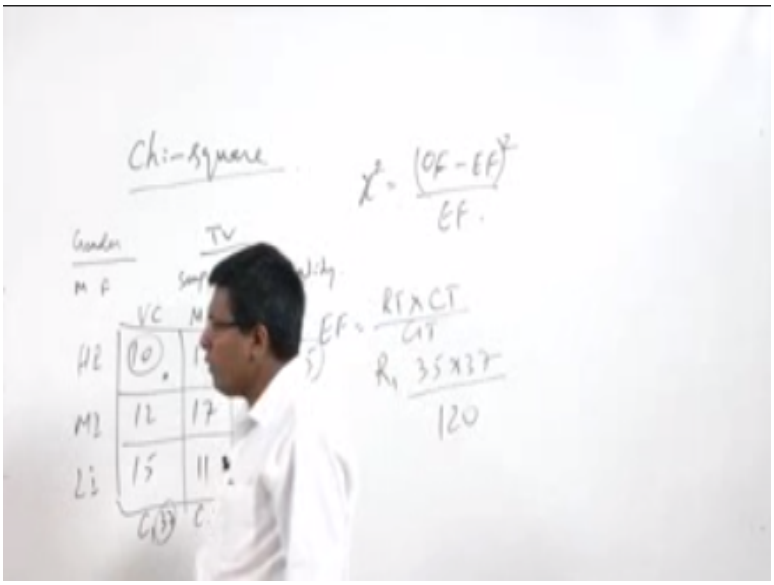
Now what does it saying, so if you look at it χ^2 is denoted like this, this symbol is equal it says observe frequency - expected frequency divided square upon expected frequency so this is the actually formula right, so okay.

(Refer Slide Time: 12:57)



So this is the χ^2 formula I have given now how do you calculated this expression frequency, now what is this let us say, let us take this example.

(Refer Slide Time: 13:06)



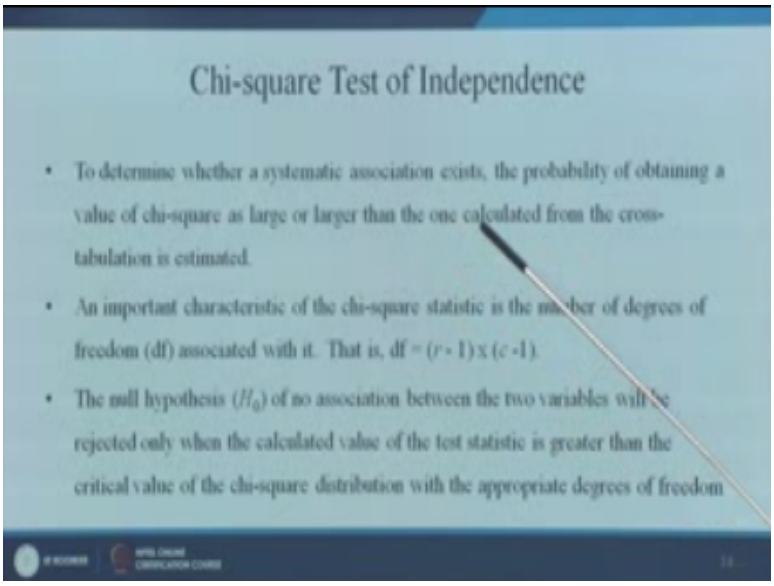
I am doing let us say I want to see okay whether there is any relationship between let us say income level and cleanliness okay, high income, middle income, low income okay and I have cleanliness very clean, moderately clean and less clean okay. Now I have got some observations I have made some observations and let us say this is 10, this is 12, this is 15, this is 12, this is 17, this is let us say 11, this is I am randomly putting numbers this is 13, this is 18, this is 12. So can I say that there is a relationship between income and cleanliness can I say that.

Now these are my observe data right, now if you see this formula I need an observe and expected frequency so what is the expected frequency that, that is what it is saying now expected frequency is equal to you have to go through the total, now 10, 20,30, 33,35 10, 20, 30, 38,45, 47 okay I hope my correction additions are correct 30, 32, 35,36, 38 okay. Similarly you can add this side so this you have to add the column total, column total right and these are the row total.

So row 2, row 3 so now expected frequency is equal to row total for that right, into for that variable let us see you are talking about this variable so R1 it will be R1 so let me write this first row total and column total by grand total okay. So in this case let us add this 10, 20, 30, 35, 37 okay this is 37 this is 10 20m30 37 38 40 okay this is 30 32 30 43 okay.

So what is my first value or excepted value now 35 x 37 right up one how much total you have to add up this three 30 40 70 30 100 108 115 120 so similarly you have o find out the expected frequency for each cell so this one so this one then next this one so for this into this divided by grand total 120 right. For similarly this you have to count okay now after doing this.

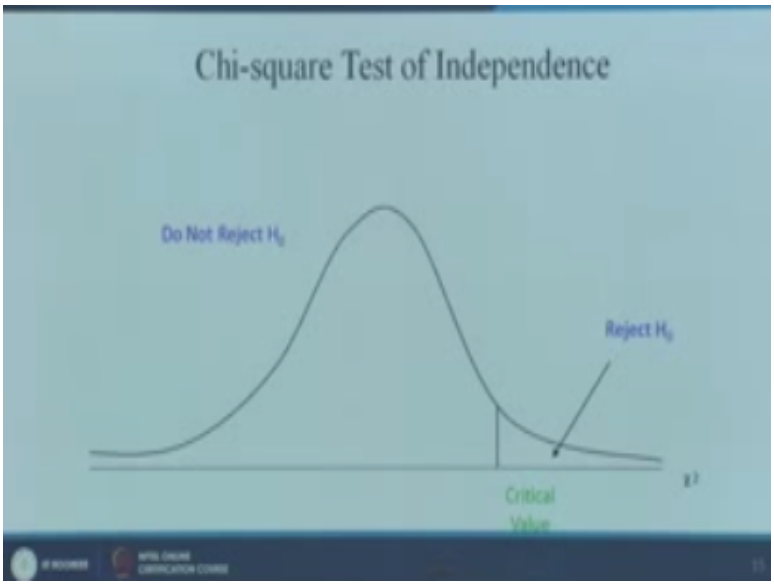
(Refer Slide Time: 15:54)



So what it says to let us say one important thing that when you do this you also have to know what is my degree of freedom because at the end of the day when you want to after you have calculated the χ^2 value the $\chi 2$ value has to be compare to the table value right for a certain degree of numbers of degrees of freedom so what is the degree of freedom, you know $\rho - 1$ x column – 1 I hope you can see this degree of freedom is equal to r – 1 x column – 1 in this case in our this case how many rows are there 3 – x 1 – of 1 x how many columns 3 so haw many so what is my total degree of freedom so this is equal to 4 2 x2 okay.

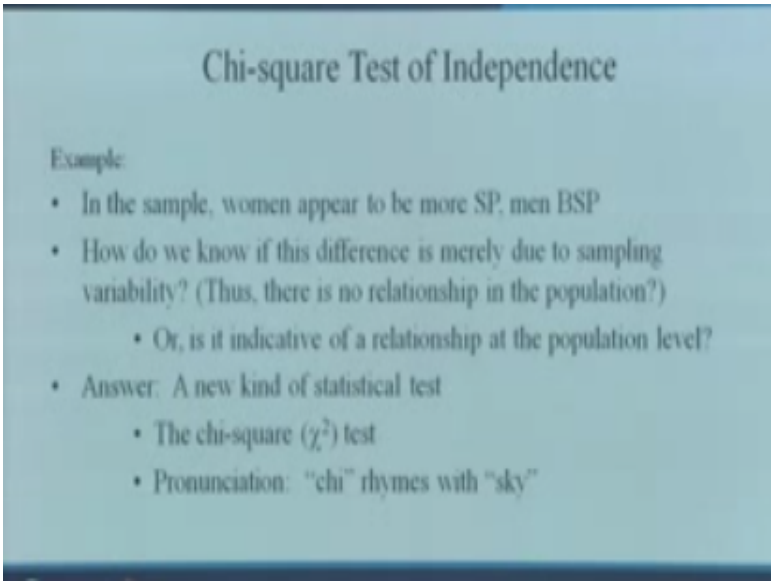
So the null hypothesis in this case is no association with the two variables is there so there is null hypothesis there is no association right, now let us see this.

(Refer Slide Time: 16:48)



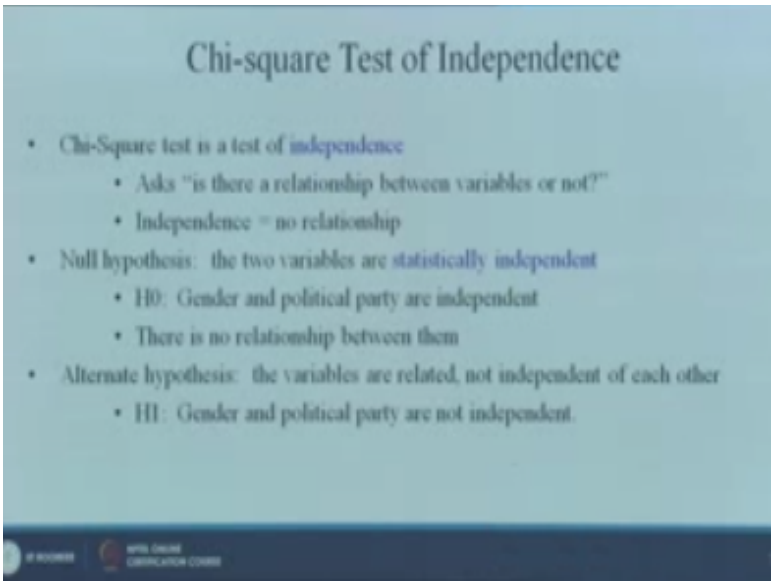
So this is exactly the chi square will not look like the chi square will looks slightly different I will show you but anyway so where does it fall we will see that.

(Refer Slide Time: 17:00)



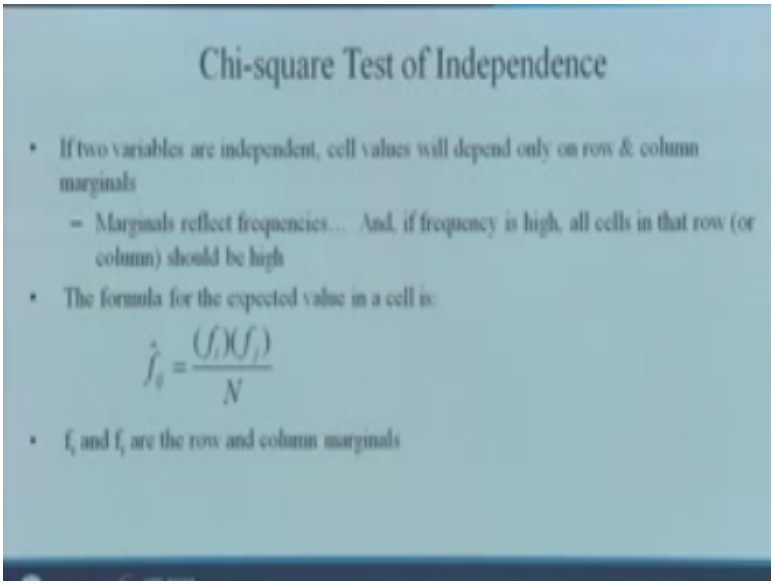
Let us say this is the example in the sample women appear to be more in SP and men with BSP I think it is what did he saying how do we know this difference is nearly due to sampling variability the question is that is this difference of men and women that you saw that we also where speaking through a cross stabilization is this just due to the sample or actually it is a way it is actually in the population also it is same thing. So let us do this to do this we say the chi square test is to be used right.

(Refer Slide Time: 17:34)



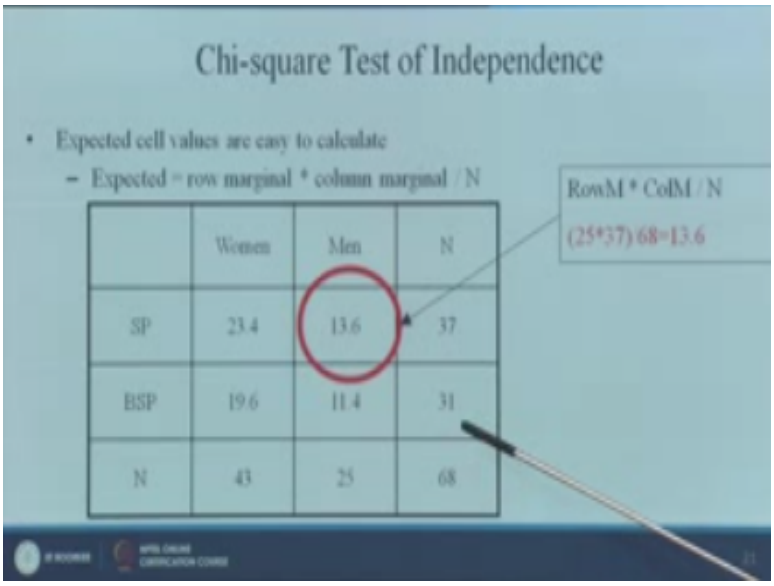
So is the test of independence it asks if there is a relationship between the variables or not right so independence means there should be no relationship the independent of each other right so what is my null hypothesis the gender and political party are independent of each other that means there is no relationship between gender and the political party in this case there is no relationship between the income and the cleanliness of people right. So alternate hypothesis is they are not independent as good as that.

(Refer Slide Time: 18:02)



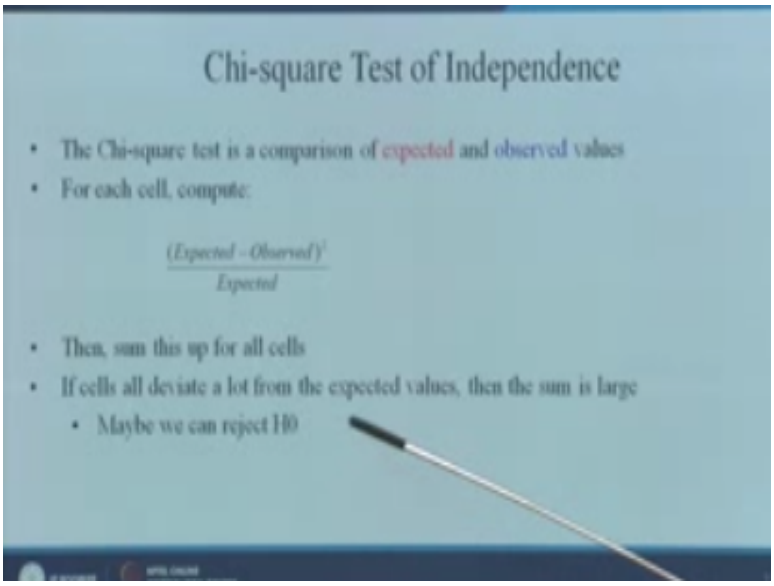
Now observed value and the expected value I was just telling you right, so expected value is as I said so you have to find out this I the value for the I that means the row total in to column total and divided by the overall total grand total okay.

(Refer Slide Time: 18:20)



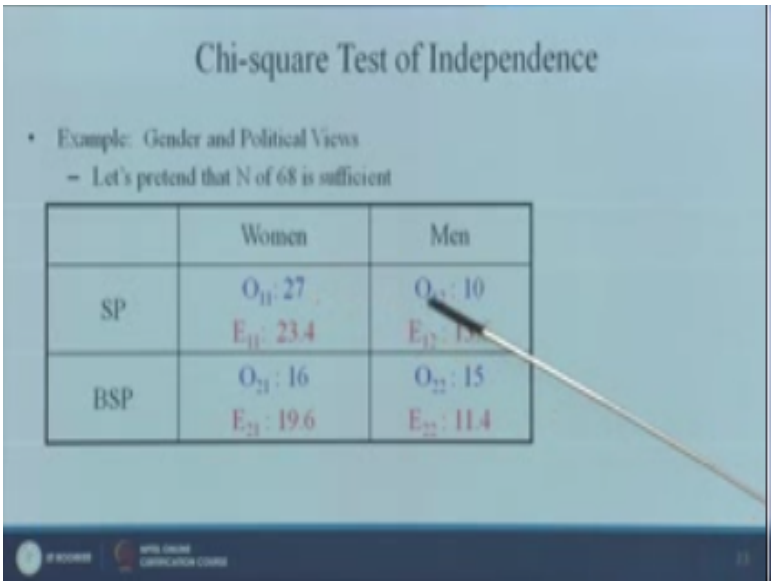
So let us say this now look at this case right so we have got this values okay now for example this one now how it is done 37 so this is 37 x 25 / 68 okay so this is 13.6 so this value for the expected one come to 13.6 okay. for let us say this one now this will be and ea h one you can see 312 x 25 so this for this one this cell is 321 the row or total in to the column total is 25 / 68 if you multiply this will come this much. So all the cells have been multiplied this is only a example which has been shown okay.

(Refer Slide Time: 19:12)



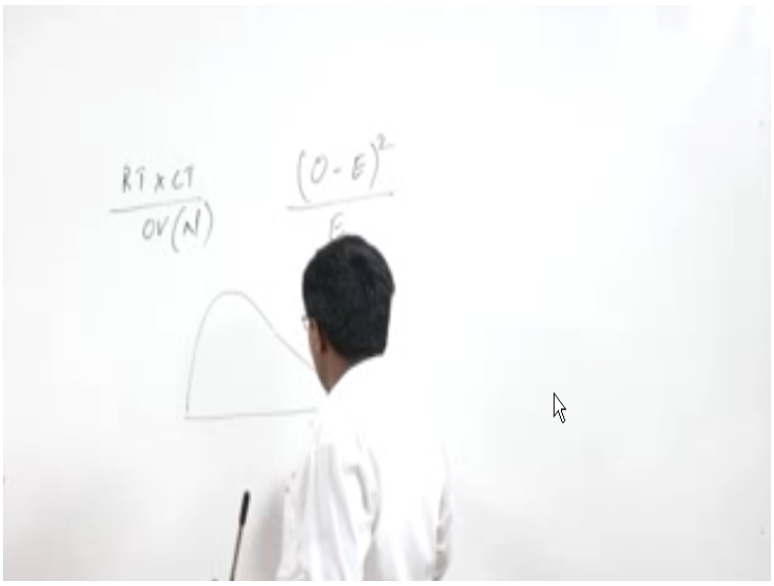
So the chi square test of independence basically it helps in comparing the expected and the observe values right so if you what it basically compute and it says the expected minus observe square upon the expected so in books will find out it is rottenly observe minus expected it does not matter because after all there is a square so the sign will not matter to us right to upon the expected okay. So in basis of this so what we do is finally we will reject our accepter our null hypothesis.

(Refer Slide Time: 19:43)



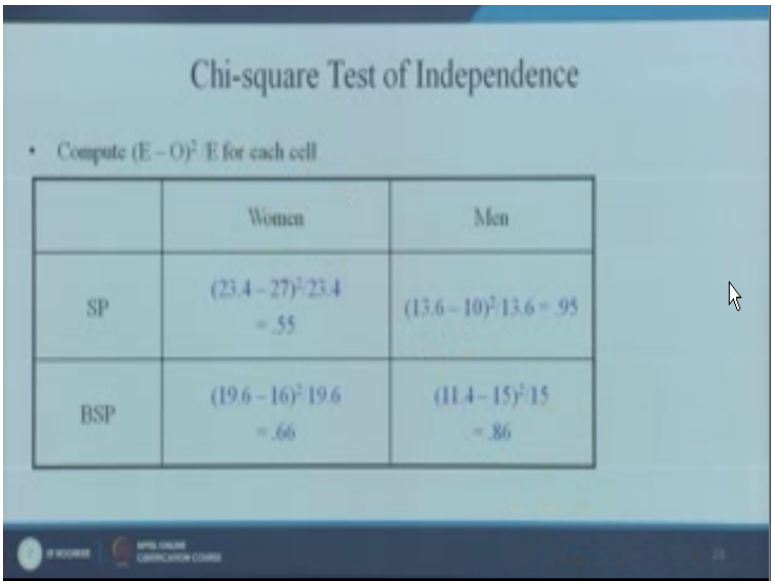
So now in this case let us say once we had the observe value of 27 10 16 and 15 right.

(Refer Slide Time: 19:56)



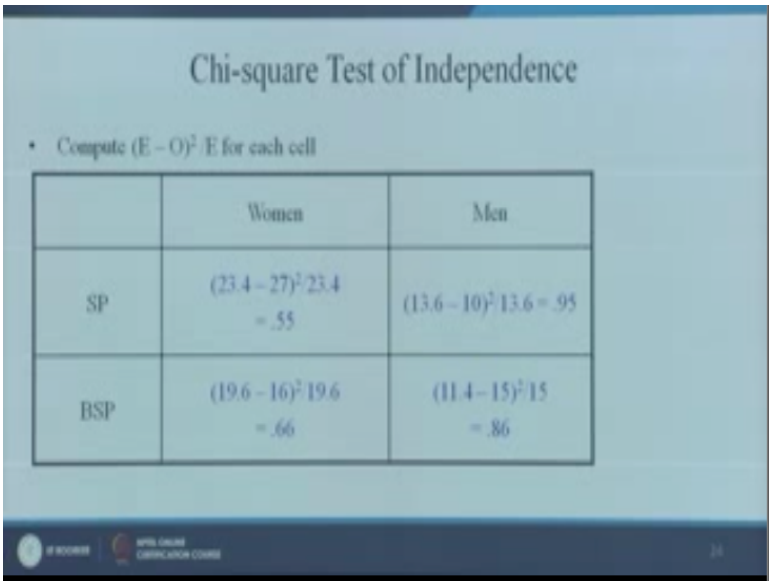
Next we calculated as I said earlier how to be calculate we did the row total*column total/ overall total or which is sometimes n also total n right so by doing this we calculated the E11=23.4, E12=13.6, E21 is 19.6 and E22 is 11.4 okay.

(Refer Slide Time: 20:22)



Now once we have calculated this from here we will calculate the expected-observed square that formal so again as I said I generally have an habit of writing observed-expected square/expected so this is important because I need the square will take care of other things right so by doing this we have found that in SP women it was the E-O^2 for E overall $\chi 2$ value came 2.55 and for man it came to 2.95 for BSP the women it came to $\chi 2$ value of .66 and this is .86 so overall $\chi 2$ value is equal to overall $\chi 2$ value is added up here.

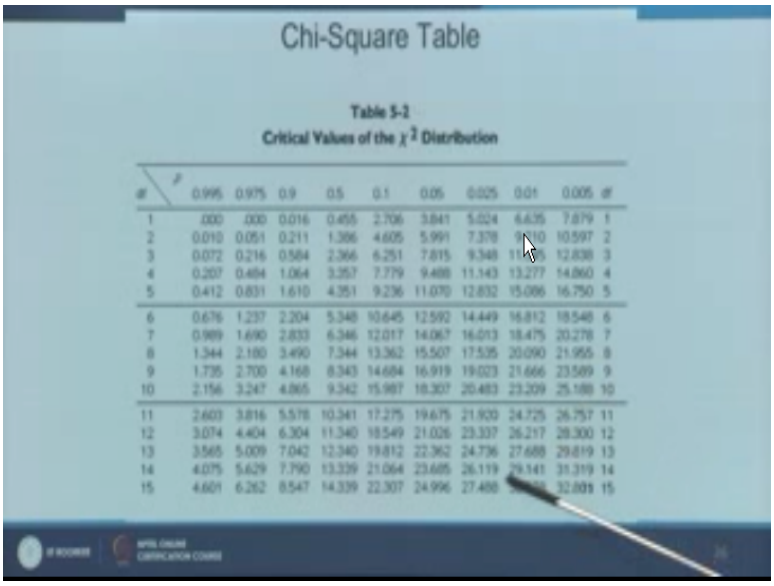
(Refer Slide Time: 20:23)



This came to 3.02 okay now what will what are we going to understand from this value right so as now we have to understand that when you have the calculated value you have to obviously compare like any hypothesis testing we have done it till now so we have to compare it again the critical value or the table value right.

So it is how it looks like so and our value is somewhere let say let us assume this is 3.02 right 3.02 so what is the critical value now so to check the critical value first you need the degrees of freedom as I said I think so degrees of freedom is calculated as R-1*C-1 so in this case we have two rows and two columns so that means 2-1*2-1 this is equal to 1 right.

(Refer Slide Time: 22:00)



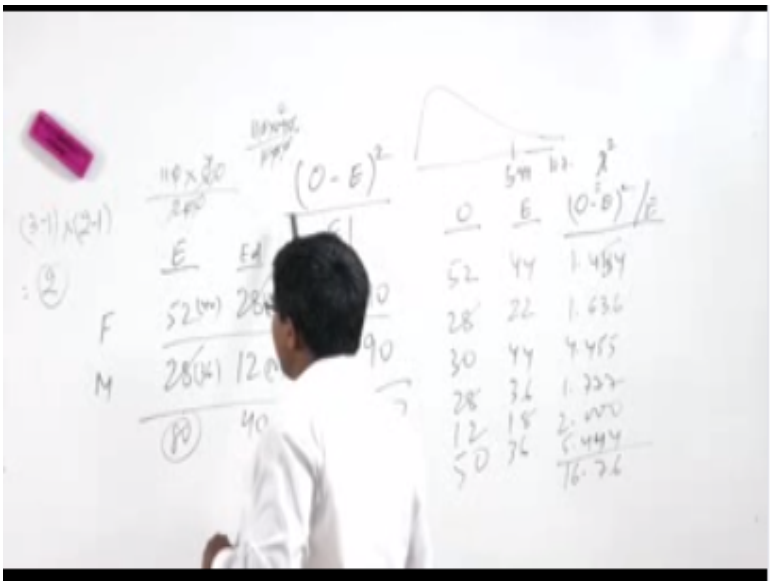
df \ p	0.995	0.975	0.9	0.5	0.1	0.05	0.025	0.01	0.005	df
1	.000	.000	0.016	0.455	2.706	3.841	5.024	6.635	7.879	1
2	0.010	0.051	0.211	1.386	4.605	5.991	7.378	9.210	10.597	2
3	0.072	0.216	0.584	2.366	6.251	7.815	9.348	11.345	12.838	3
4	0.207	0.484	1.064	3.357	7.779	9.488	11.143	13.277	14.860	4
5	0.412	0.831	1.610	4.351	9.236	11.070	12.832	15.086	16.750	5
6	0.676	1.237	2.204	5.348	10.645	12.592	14.449	16.812	18.548	6
7	0.989	1.690	2.833	6.346	12.017	14.067	16.013	18.475	20.278	7
8	1.344	2.180	3.490	7.344	13.362	15.507	17.535	20.090	21.955	8
9	1.735	2.700	4.168	8.343	14.684	16.919	19.023	21.666	23.589	9
10	2.156	3.247	4.865	9.342	15.987	18.307	20.483	23.209	25.188	10
11	2.603	3.816	5.578	10.341	17.275	19.675	21.920	24.725	26.757	11
12	3.074	4.404	6.304	11.340	18.549	21.026	23.337	26.217	28.300	12
13	3.565	5.009	7.042	12.340	19.812	22.362	24.736	27.688	29.819	13
14	4.075	5.629	7.790	13.339	21.064	23.685	26.119	29.141	31.319	14
15	4.601	6.262	8.547	14.339	22.307	24.996	27.488	[illegible]	32.801	15

So the critical value let us go at 5% level so if you look at this able and then you go to .05 5% is what you wanted right so when p value is .05 right and you have degrees of freedom is only 1 so you have a 3.841 correct so now you can compare it again this and you can find out whether you have to expect or reject.

So our value was lees 3.02 so that means we cannot conclude at this moment that there is a relationship between gender and political party of relation let us have a problem which I had brought to for you will do this so this problem is something like this right it says there are 200 people okay random is selected adults there are asked about the TV shows their opinion right.

Both male and female were ask you w hat do they feel about the TV shows right whether the TV shows are only having an entertainment value they are having a educational value or it is just a waste of time okay so the you know broadcaster they wanted to find out whether there is any relationship exist between to suggest there is a relationship between gender and the TV opinion about the TV serials.

(Refer Slide Time: 23:38)



So to do this let us solve this problem out here right you can continue with me side by side when I am giving the data so let us say they were female and male right now who thought entertaining is of this entertaining some thought this educational some thought this is a waste of time so 52 women thought this is a entertaining 28 means thought it is a entertaining 28 thought education women thought.

And male 12 male thought this is an educational in waste of time 30 women thought this is a waste of time 50 men thought it is a waste of time right so now let sun see if let us solve this so in a first step what we have to do is we have find the column total and row total let us see with the column total so the column total I have brought it already so 80 this is 40 this is 80 okay and this side if we go to the row totals it is 110 this is 90 okay.

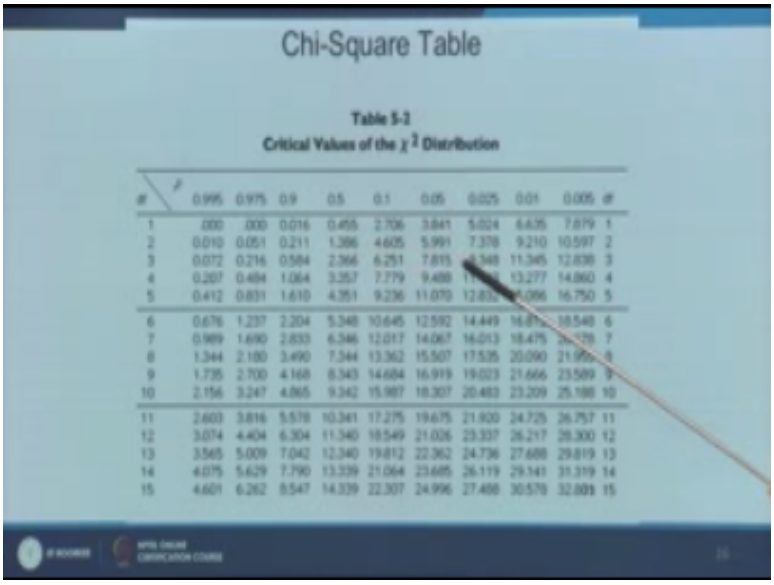
So the overall total which you are said as n basically is how much 110+90= 200 right so you go in this way or that way right so now let us calculate let us who do we calculate we have an observed values we have the expected values with us now let us find the observed- expected square right upon the expected so this are 9 χ2so χ2 is equal to this right so when I am writing this 52, 28,30, 28 , 12 and 50 okay, 44, we calculated this.

So have you calculated no we have not calculated, I have brought so I am saying, so let us calculate the expected values. So for this one 110, let say for this one 110 x how much 80 okay upon the grand total which was 200, so this was 2, 4, 11 for 44, so this comes 44 okay, similarly for this one 110 x 40 so upon 200, so 200, 2, 2 ,2 so 22.

So similarly we can calculate for others right, so when I am calculating for others I have already done it, so I am writing it here 44 this is 36, this is 18 and finally this is 36. So this value only now I am writing out here, so 22, 44 okay 36, 18 and 36 and from here let us finally calculate χ^2. So 1.454 sorry this is 1.636 right, 4.455, 1.777, 2.000 and 5.444. So if we take the overall total it is something around 16.76 okay.

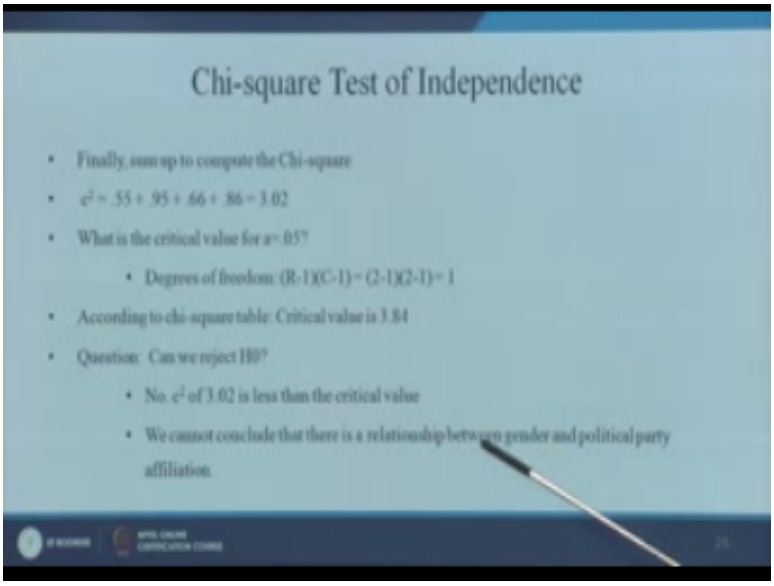
Now this is our total, now we have calculate against the table value, so the table value critical value is now we need to find regressive freedom,. Now there are three columns, so three columns – 1 x 2 ρ – 1, so that means what we have got two rights. So we will have to find the value for two degrees of freedom right and let us check how much it is coming.

(Refer Slide Time: 27:24)



So for degrees of freedom at a 5% level of significant the value is how much, now 2^0 okay 5%, so 5.991. So if it 5.991, so is the cut of volume 5.991 and we have got the value of 16.7 which is beyond the critical value right. So that means what our null hypothesis is rejected and what are our null hypothesis is that the gender and the opinion about the programs is indep0endent of each other right. So if you go back let see this.

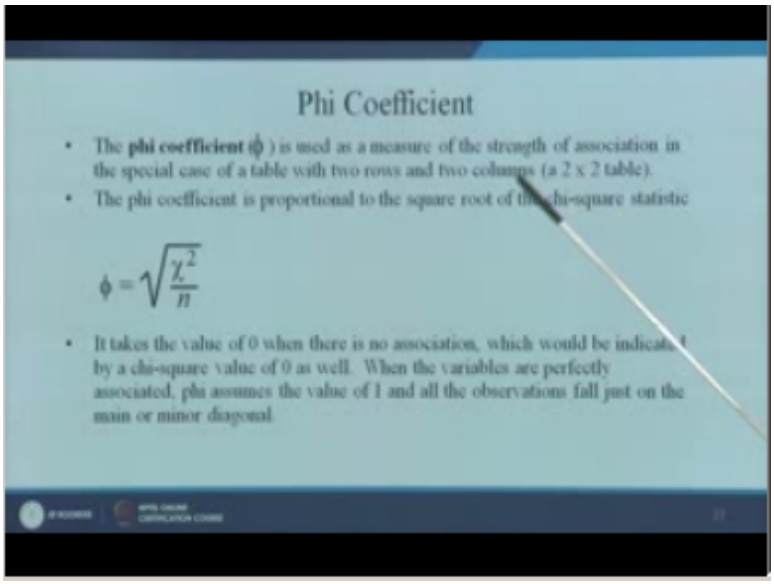
(Refer Slide Time: 28:15)



As you had here gender and political party affiliation right, there is no relationship between the gender and political party affiliation similarly in this case it is becoming though the relationship between the gender and the opinion about the programs right. But in this case we could not conclude, the null hypothesis had the excepted but in this case which we have solved here, we are saying that the null hypothesis is rejected that means, there is the significant difference what female feels and what males feel.

So this how you calculate the chi square value and you test and hypothesis to the χ^2 right. Two things I am going to show you here besides this, these are the some other values which you would see.

(Refer Slide Time: 29:02)



This is called the ϕ coefficient which is used in case of any software if you are using any special cases or something you will see this and sometimes people get confused to understand what it actually implies right so the ϕ coefficient is nothing but it is the $\sqrt{\chi^2/n}$ right so what did basically does is helps in measuring the strength of association between two variables right so between two variables like in any test right.

So the ϕ coefficient when it is closed to 1 when it is 0 there is no association and when it is closed to 1 we will feel they are perfectly associated there is a high degree of association between the two variables okay one more data that you see is called the contingency coefficient right so this is shown as c which is nothing $\sqrt{\chi^2/\chi^2+n}$ right so it varies between 0 and 1 the maximum value of the contingency coefficient why I am saying this is you might require it to write down during the research paper it depends on the size of the table.

Okay the number of rows and number columns right because that is the end basically right for this reason it is to be used only to compare tables of the same size if you have tables of different sizes then you should argue you cannot compare it right so this is one thing and one last thing which I have forgotten is you to remember that when you do a χ^2 test please remember that the cell values.

(Refer Slide Time: 30:38)

Have to be at least 5 or more than 5 right otherwise it does not give a it is not wise to use χ^2 test so this is all about the cross tabulation and χ^2 test and I hope you have understood why it is very simple but very powerful technique which is used to used when you have nominal data and you have any question that you collect or something you have lots of nominal data right for example gender religion sometimes suppose income for example let us say you have income is not a categorical data income is a continuous data.

But you can convert that continuous data into a categorical data now how it is for example let us say in this case for example let us you have people of income let say if somebody says my income in 2,60000 somebody says mine is 3400 let us say forget the zeros 2 lakhs right somebody says 5lakhs now what you can do is you can say anybody in-between one to three lacks is only one anybody would 3 to 5 is 2 and anybody above 5 is 3 so now we have converted the continuous data into a categorical data right.

And then again you can use χ^2 but if you convert a continues data to χ^2 there is no doubt that there will be a loss of information right but you can do it right if required if you want your research wants to get an idea how people are behaving in particular income groups so if you want to take continuous all the data it might be duff so it is better to convert it into a categorical data and do a χ^2 test and see whether there is a relationship between may be like income.

And as I was showing earlier but with my example income and cleanliness and let us say religion and spending habits so these are the things which you can do through a χ^2 test so I hope you have

understood your it is very powerful technique which you can always use and whenever in question you have demonstration variables you are free to use this techniques and you can get wonderful results out of it which I will be highly useful thanks for this session will meet in the next session thank you very much.

For Further Details Contact

Coordinator. Educational Technology Cell
Indian Institute of Technology Roorkee
Roorkee 247 667
E-Mail Etcellitrke@gmail.com etcell@itr.ernet.in
Website: www.itr.ac.in/centers/ETC. www.nptel.ac.in

Production Team
Sarath Koovery
Mohan Raj. S
Jithin. K
Pankaj saini
Graphics
Binoy. V.P
Camera
Arun. S

Online Editing
Arun.S
Video Editing
Arun.S

NPTEL Coordinator
Prof B.K Gandhi

An Educational Technology Cell
IIT Roorkee Production