Business Statistics
Prof. M. K. Barua
Department of Management Studies
Indian Institute of Technology – Roorkee

Lecture - 55
Chi-Square Test of Independence

I welcome you all in this session. As you are aware in previous session we were discussing about chi-square goodness of fit test wherein we have seen how observed data and expected data are similar to each other. So in chi-square goodness of fit test we see the expected frequencies and our actual frequencies whether they are similar or not. If they are not similar, then we sometimes reject the null hypothesis in goodness of fit is that they are similar to each other right.

So if they are not then we reject the null hypothesis otherwise we do not reject the null hypothesis. We have also seen chi-square goodness of fit as a replacement of Z test for hypothesis testing of proportion of one sample and in fact we did see how to apply chi-square goodness of fit test for comparing proportions of 2 groups or 2 populations. In previous class we did discuss hypothesis testing of proportions of 3 groups.

And we did solve this particular question using chi-square goodness of fit test. So we will see once again I will quickly go through this question.

(Refer Slide Time: 02:12)

In an area there are three hotels, the table shows response to a question, whether they will select it again

Choice of hotel	Golden palm	Palm Royale	Palm princes	Total
Yes	128	199	186	513
No	88	33	66	187
	216	232	252	700

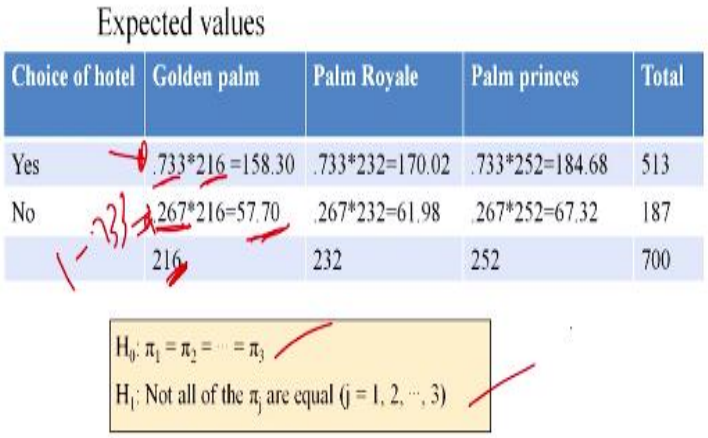
$$\bar{p} = \frac{X_1 + X_2 + \Lambda + X_c}{n_1 + n_2 + \Lambda + n_c} = \frac{X}{n}$$

(128+199+186) / (216+232+252) = 513/ 700 = 0.733

So the null hypothesis here is that the proportions are same right. So π1, π2 and π3 are same

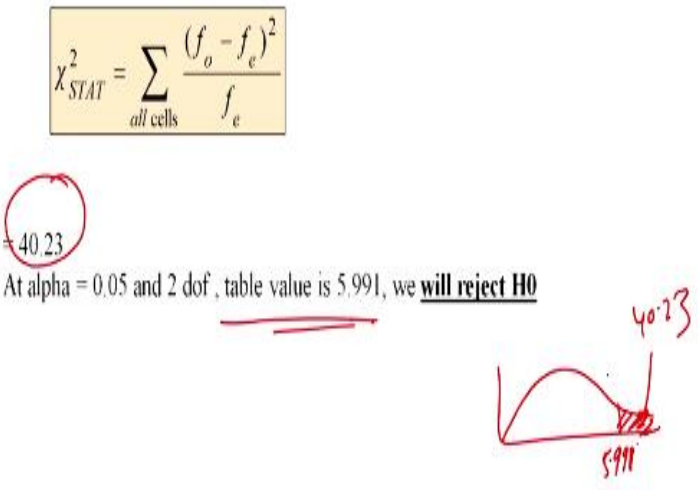
right and one of them is different from others right is alternative hypothesis. So we need to calculate this overall proportion first of all which is 0.733 so all these observed frequencies right.

(Refer Slide Time: 02:45)



And the expected frequencies would be 0.733 multiplied by 216 for those customers who would come again to this particular hotel and 0.267 how did we get this 1- 0.733 right is 0.267 multiplied by 216. So 57.77 70 customers said they will not come to that particular hotel right. So this was our null and alternative hypothesis.

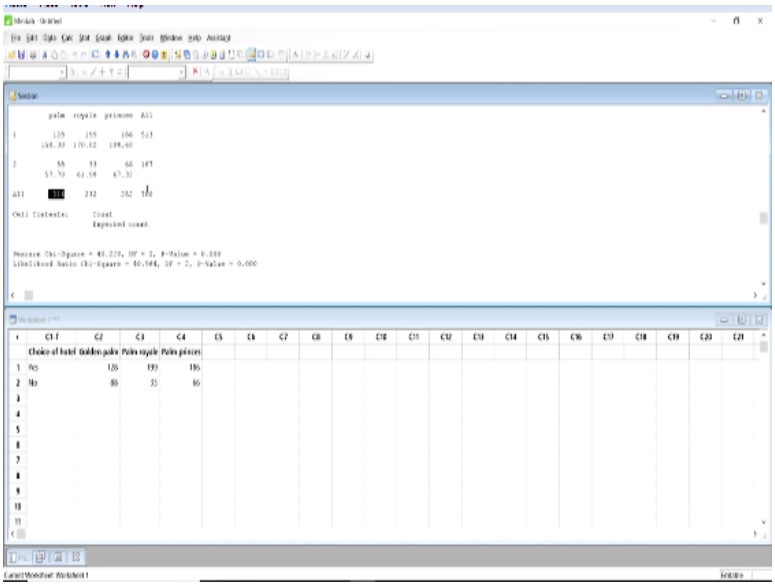
(Refer Slide Time: 03:21)



And when we calculated this chi-square value it came out to be 40.23 and the critical table value was 5.99 so we rejected null hypothesis so this was our rejection region critical value 5.991 right and this was our calculated value 40.23 right. So we will reject the null hypothesis

and we will say that these 3 proportions are different right. So let us solve this question using Minitab.

(Refer Slide Time: 04:02)



So first of all we will go for data entry so choice of hotel so you have got golden palm, palm royale and palm princes. Now the choice of hotel yes or no. So these are observed frequencies. There is no need to write total values for these 2 rows and there is no need to write even total of column values right. So just go to stat, tables, chi-square test of association in fact you can apply this.

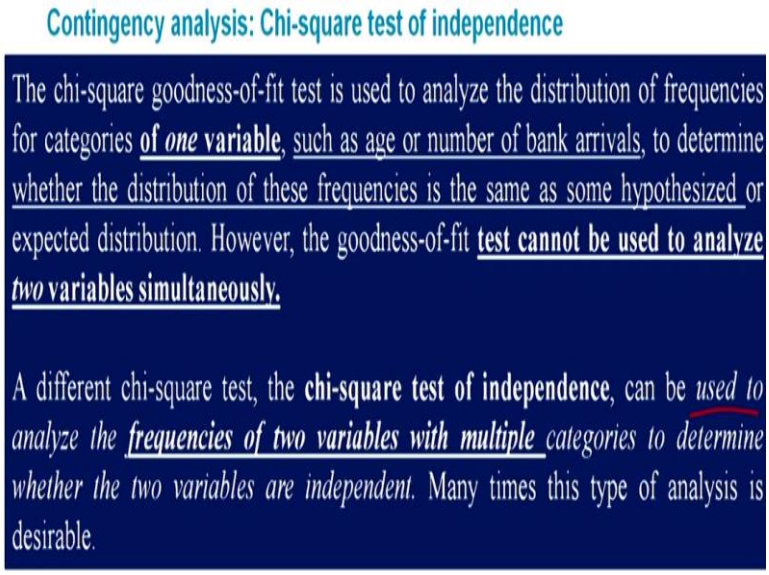
So select all these 3 statistics. So we have got the we are interested in chi-square statistics which we have clicked okay right. So just look at this these are in fact expected value right is not it and how did we calculate these expected value in our question. It was just row total it means 513 * 216/700 right. It was row total * column total/total number of observations right.

So let us look at chi-square value first look at P value P is 0 right. So P< alpha so we will reject the null hypothesis and the null hypothesis was that all these 3 proportions are same. So we are saying they are not same because we have rejected null hypothesis. So for a chi-square value calculated is concerned it comes out to be 40.96 just check whether we were getting the same answer yeah 40.23 so approximately one and the same thing, in fact, yeah this is exactly same right.

So this is Pearson's chi square is 40.228 just see this right. This is what we have calculated

over here 40.23 okay. So we will reject the null hypothesis.

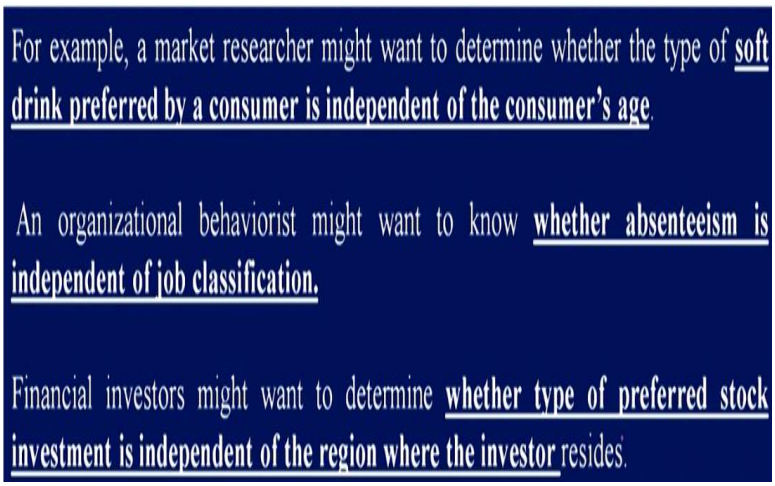
(Refer Slide Time: 06:57)



Now let us look at some more applications of chi square test. So we have seen chi-square goodness of fit test now let us look at chi-square test of independence. Generally, chi-square goodness of fit test is used to analyze distribution of frequencies for categories of one variable. So the examples which we have seen so far they were having only just one variable. For example, the numbers of customers arriving at a bank counter, determining the distribution of whether the frequencies are same as hypothesized or as expected.

So the goodness of fit test cannot be applied when you need to analyze 2 variable simultaneously right. So for that you need another techniques called chi-square test of independence. So this can be used to apply to analyze frequencies of 2 variables with multiple categories to determine whether 2 variables are independent or not.

(Refer Slide Time: 08:12)

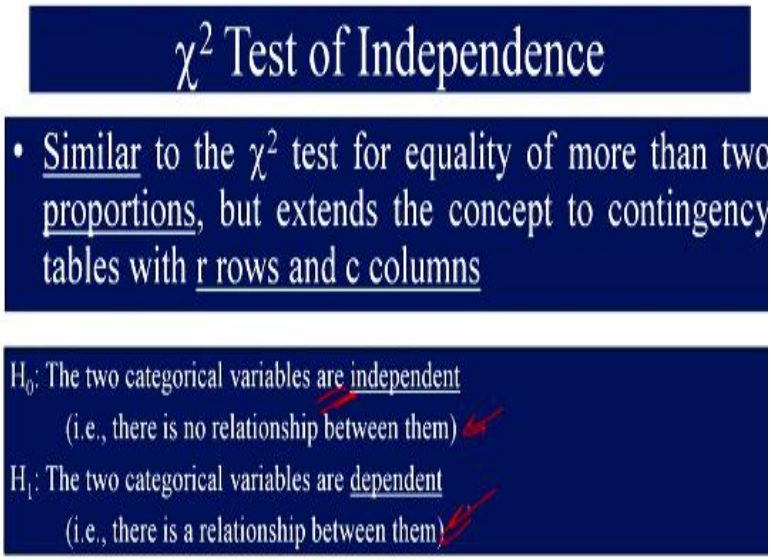


So let us look at what could be the examples of chi-square test of independence. So let us say we want to know whether the soft drink whether the type of soft drink preferred by consumers is independent of consumer size. So are they dependent or they are independent right. So we always say there are independent right. For example, in organizational behaviorist might want to know whether absenteeism is independent of job classification or whether the type of preferred stock investment is independent of region where the investors resides.

So it is the test of independence. So we will say whether age is independence of coffee consumption. Whether studies is independent of good score whether more studies independent of good score and so on right. So you can have several types of hypothesis tests wherein the 2 variables are independent of one another. So this is very similar to chi-square goodness of fit test.

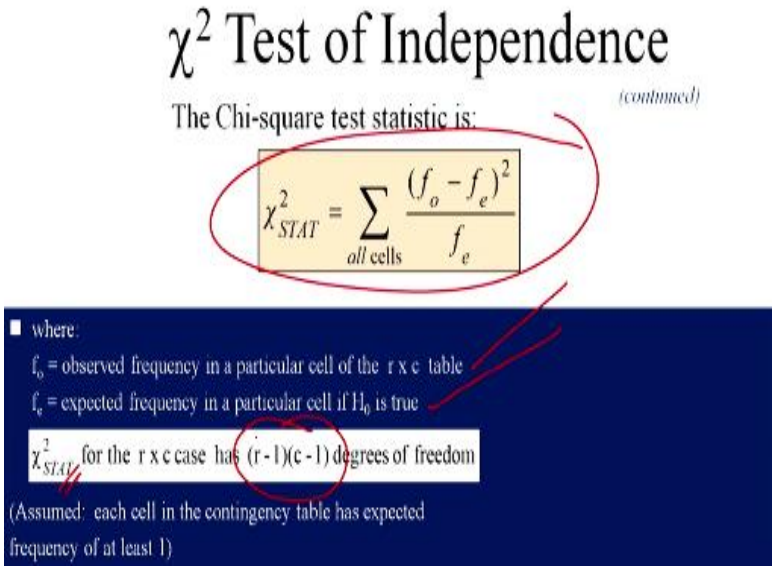
And this is similar to what we have seen earlier wherein we did compare proportions of more than 2 groups right. So the same concept is being extended over here.

(Refer Slide Time: 09:57)



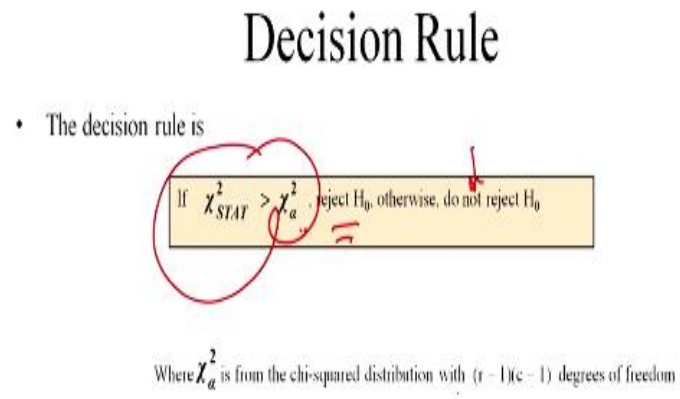
So null hypothesis is 2 categorical variables are independent right. So there is no relationship between them and alternative is there is relationship between them right.

(Refer Slide Time: 10:11)



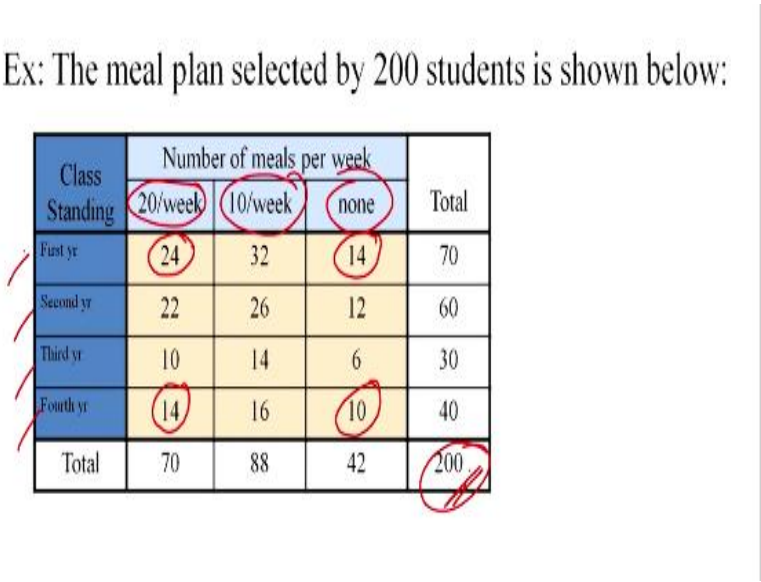
So how to calculate chi-square statistics same formula here say fo- fe whole square divided by fe. So fo observed frequencies, fe expected frequencies chi-square statistics value the critical values used to be seen at r-1 and c-1 degrees of freedom right. So here R is number of rows C is number of columns.

(Refer Slide Time: 10:40)



Decision rule remains same there is no change in it if chi-square statistics is more than critical value then we will reject the null hypothesis otherwise not okay and at this particular degrees of freedom right.

(Refer Slide Time: 10:59)

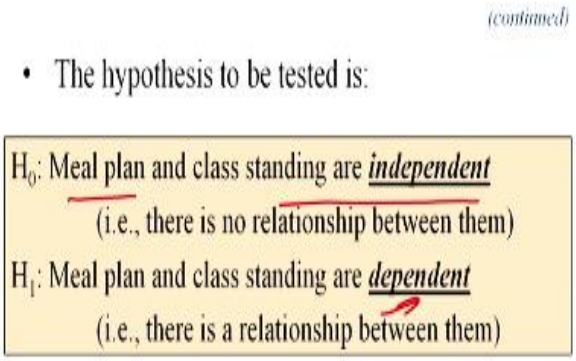


So let us look at this question the meal plans selected by 200 students in an institute is shown in this table. So these are total 200 students right. So the class standing can be in first year, second year, third year and final year right or fourth year. The number of meals per week the options available are 20 meals per week 10 meals per week and none per week. So we will say that there are 24 first year students who have taken 20 meals per week, 14 did not take even a single meal per week in a mass let us for simplicity.

Similarly, 14 students of fourth year have taken 20 meals per week and 10 of them did not

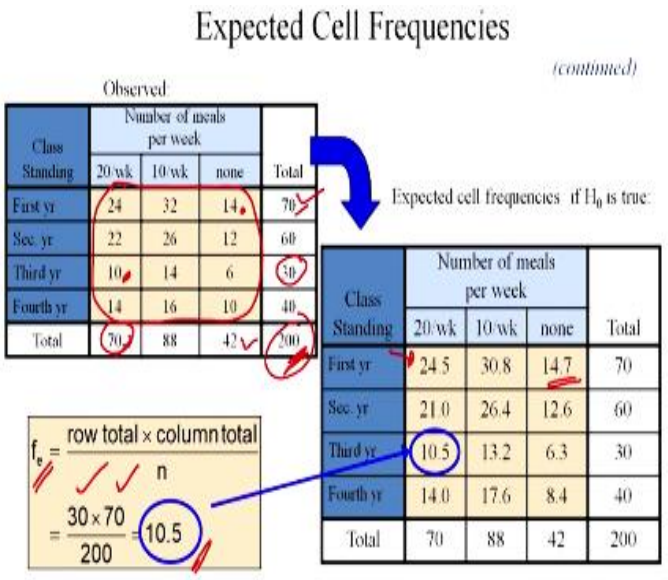
take even a single meal per week. So the null hypothesis is class standing independent of number of meals per week or in other words is there any relationship between the number of meals taken by students and their class standing. So this is to be checked right.

(Refer Slide Time: 12:31)



So we will say that null hypothesis meal plan and class standing are independent right and their dependence is alternative hypothesis right. So meal plan and class standing are independent or there is no relationship between them is alternate is null hypothesis.

(Refer Slide Time: 12:54)

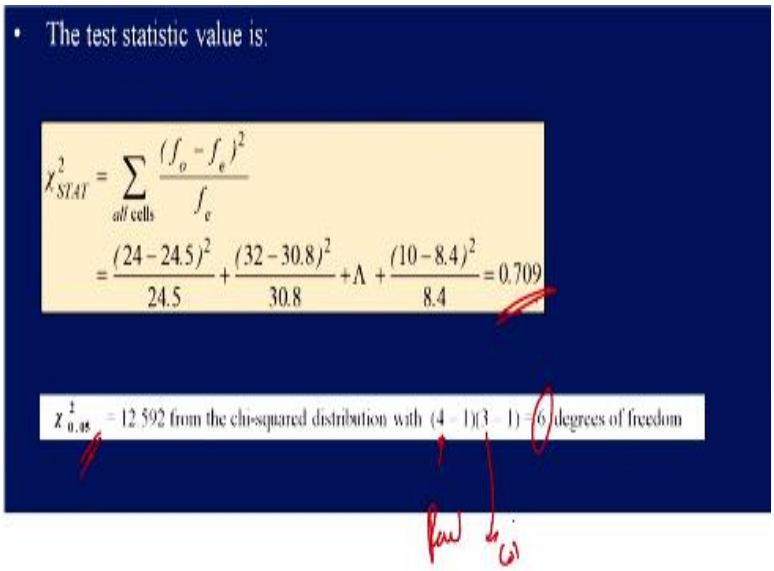


So first of all you need to calculate expected value for each of these cells these are different cells wherein we have written observed frequencies right. So fe are expected frequencies row total/column total so 70* 70 right/200 you will get 24.5 right. Let us say if you want expected frequency for this particular cell so 70* 42/200 we will get this. Similarly, for this what

would be the expected frequency row total which is 30* column total which is 70 right multiplied by multiply both of them and divide it by 200 right so this is 10.5.

So this is how you can calculate expected frequencies.

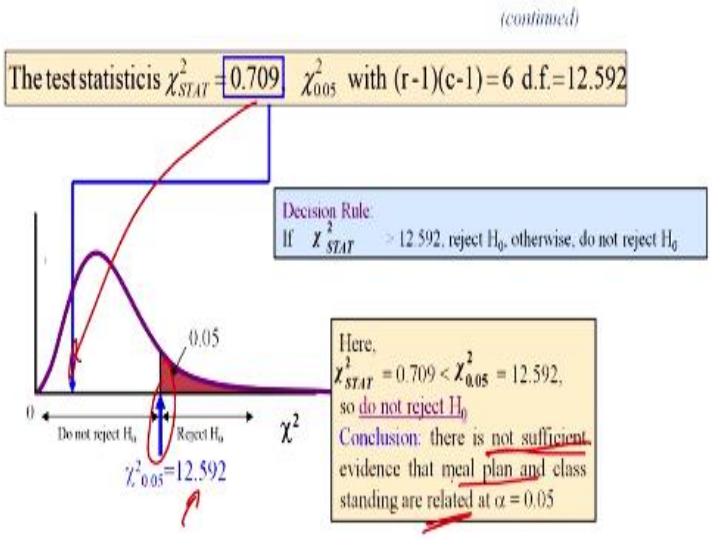
(Refer Slide Time: 14:01)



When you solve this question by chi-square statistics this comes out to be 0.709 and at alpha-0.05 and at 6 degrees of freedom. So it is number of rows and C is number of columns right. Let us see how many rows and columns are there in this question. So there are 4 rows 1, 2, 3, 4 and columns are 1, 2 and 3 right. So row-1* column-1 will be degrees of freedom. So the critical value is this and observed value is this.

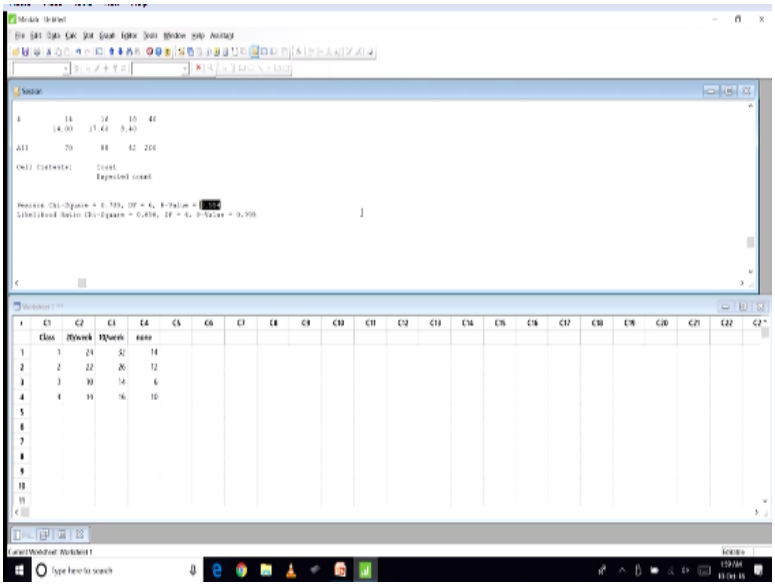
So in fact you can check whether this is chi-square statistics is this. So at 6 degrees of freedom this is the one and at 95 right this one 12.592 right so this is 12.592 is not it this is how you can check table value of chi-square.

(Refer Slide Time: 15:14)



So this one is your critical value it is 12.592 and the calculated statistics is this which is in non rejection region is not it. So we will not reject null hypothesis. We will not reject null hypothesis, it means they are independent right. So there is no sufficient evidence that the meal plan class standard are related. So they are not related with each other they are unrelated with each other okay.

(Refer Slide Time: 15:53)

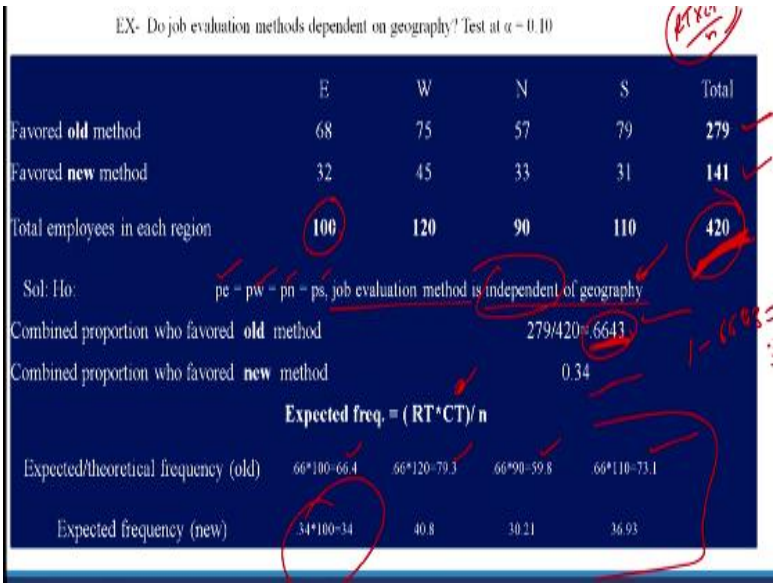


So let us look at how to solve this question using Minitab. So first of all you just delete all these output and we will have data entry once again. So class standing first, second, third and fourth here right. You have got different meal plans 20 per week, 10 per week and none per week. So these values are 24, 32, 14 then you have 22, 26 and 12, 10, 14, 6, 14, 16 and 10. Let us solve this question.

So table and in fact you can directly go for chi-square test for association is not it. So this is what you have obtained so the chi square is this just 0.709 exactly same as what we were getting right 0.709 is not it and the expected values can also be seen over here. So let us say this is the expected value for so for first year expected value is this right. So it is 24.50 how did we get this.

So row total * column total/200 right. So we are getting same answer right. So this is how you can look at other values you can just go down yeah just look at the P value is 0.990 right while alpha was 0.05 right. So P value is less than alpha no. So we will not reject the null hypothesis right that is what we did in our question right. So this is how you can use Minitab to solve questions on chi-square test of independence. Let us look at some more examples.

(Refer Slide Time: 18:44)



So the question is like this a survey was conducted amongst nurses of hospital and the survey was conducted in 4 different regions of a country east, west, north and south and they have asked the question whether they would like to have salary every year which is the old method every month or they would like to have salary twice in a month. So when the question was asked to let us say 100 nurses of east, 68 favored the old method it means they wanted salary every month.

While 32 said that they wanted twice a month right. Similarly, for west, north and south right. Now we need to test this hypothesis that that 0.10 significance level. So the question is are these 4 proportions same or we can put it like this. The proportion of nurses preferring new methods is same across these regions is not it. So you need to have null and alternative

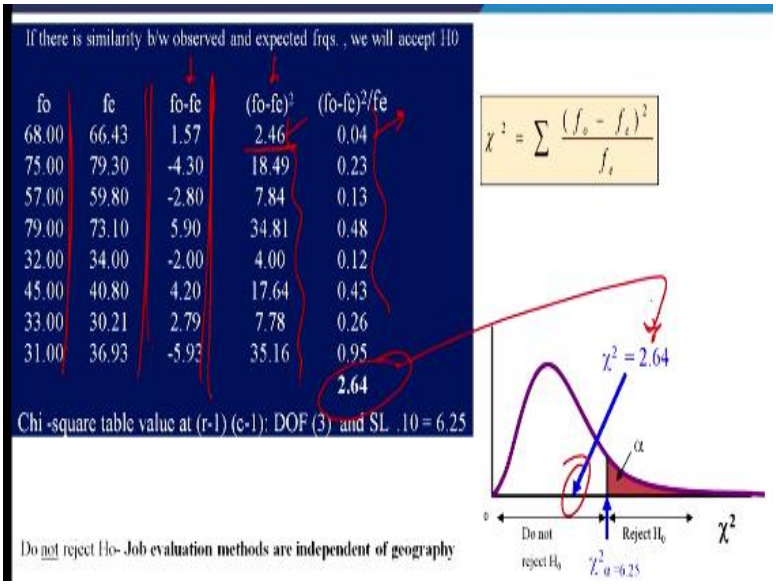
hypothesis. So you have contacted total 420 nurses, 279 preferred old method, 141 preferred new method.

So null hypothesis is this proportion of nurses are let us say the job evaluation method is independent of geography right. So you can frame your null hypothesis in this way. So we are saying that each job evaluation method. In fact, job evaluation is nothing, but how are they performing right. So proportion of nurses in east, west, north and south is same is our null hypothesis right.

So alternative hypothesis would be is dependent only this much change will be there. So the expected frequencies is again row total * column total/420 right. So or you can have expected frequencies in this manner as well. You just calculate the overall proportion first of all so 0.66*100 is this 0.66 multiplied by 120 is this 0.66 multiplied by 90 is this 0.66 multiplied by 110 is this right.

And how to get this one it is 1-0.663 right .6643= 0.34. So 0.34 multiplied by 100, 120, 90 and 110 right. So either you calculate expected frequencies using this way or simply row total * column total/total number of observations right. So I think the method of calculating expected frequencies row total * column total/ n is simple right compared to this one okay.

(Refer Slide Time: 22:35)



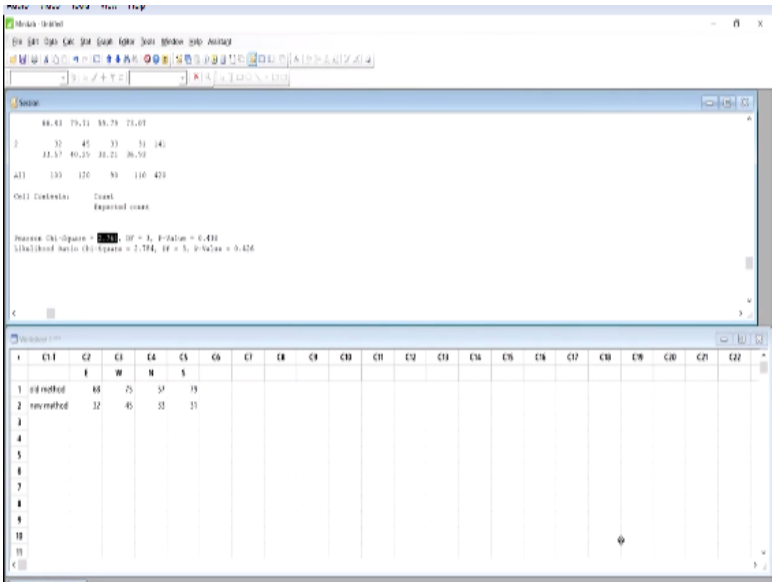
Let us look at the solution to this question so you have got all this fo over here all fe over here, fo-fe is some of them are positive, some of them are negative. The moment you take square of this particular column which is here in this column all values are positive divide

each of these values/fe right. So let us look at this 2.46/ 66.43 right so you will get 0.04 and so on.

So when you take the sum of this it is 2.64 so calculated chi square statistics is 2.64 which is this right. Now you need to look at the critical value chi-square critical value from table at appropriate degrees of freedom so how many rows and how many columns. So there are 2 rows so row-1 right that is 2-1 multiplied by column-1 columns are 4 right. It means so 4-1 so this is 1 and 3 so this 3 degrees of freedom right so just see.

The critical value at 0.10 significant level at 3 degrees of freedom is 6.25 which is here so this is your critical limit calculated value is in non rejection area right. So we will not reject null hypothesis right. So we will not reject null hypothesis it means job evaluation method are independent of geography. So there is no relationship between job evaluation method and geography. In other words, we will say they are independent right. They are not dependent on each other.

(Refer Slide Time: 24:46)



So let us look at this equation and we will solve this question using Minitab. So east, west, north and south old method, new method so you just enter data in fact you need not write total values row of total and row of columns just multiply just enter data go to stat chi-square test for association. Just select all these 4 values click okay right. So let us look at chi-square yeah it is 2.76 is Pearson chi-square so 2.76 is what we are getting.

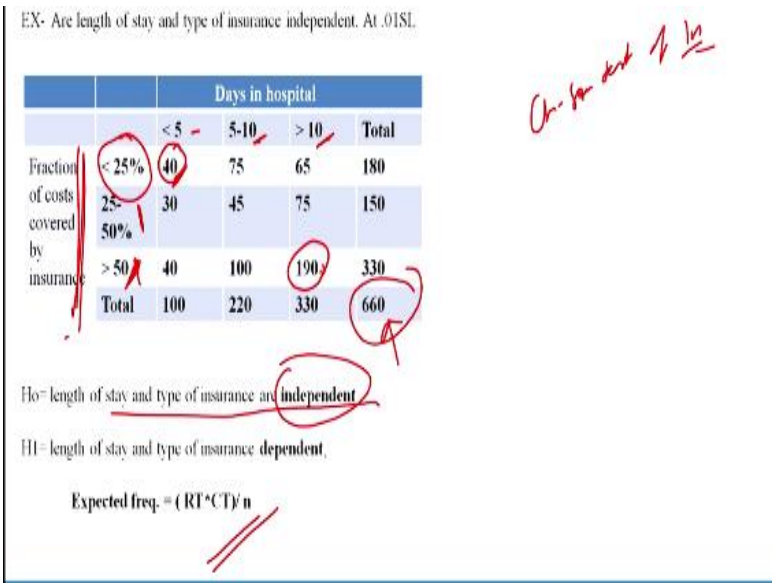
So this is the calculated chi-square value and the critical limit is 6.25 right. So let us look at P

value whether we are rejecting or not right. So P value is 0.430 is this value less than alpha. Let us look at this so what is alpha over here alpha is 0.10 right 0.10 is alpha right. And P value is 0.43 is P value less than alpha. So we will not reject the null hypothesis this is what we are saying here as well right is not it.

So you need to you can in fact collaborate your output of Minitab with the manual calculation which you have done right. Let us look at one more question on this and then we will wind up today session. So you would have observed that those people who have got insurance policy whenever they go to any hospital for treatment. In fact, the hospital keeps them for longer period of duration in the hospital, they generally do not discharge patient early.

So a set of data was collected and the objective was to know whether is there any relationship between length of stay and type of insurance whether people have got insurance or tend to stay more in hospital so that is the questionnaire that was the null hypothesis.

(Refer Slide Time: 28:02)



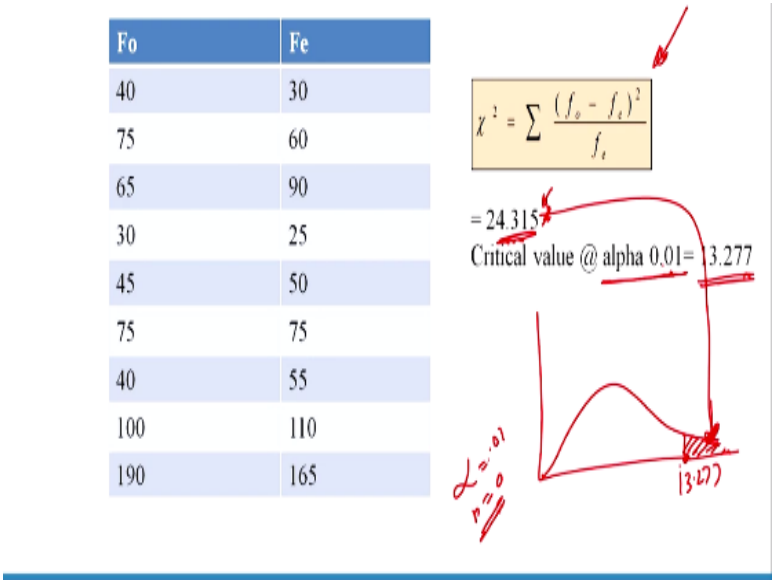
So let us look at this question. So the fraction of cost covered by insurance company if it is let us say less than 25% of the total cost of the let us say total bill of the hospital. When the cost of covered is between 25% to 50% sorry yeah 50% and when it is more than 50%. So and the stay in hospital is less than 5 days between 5 to 10 days and more than 10 days right. So let us look at what is this.

So the length of stay is when the fraction covered is less than 25% there were 40 patients and the days' number of days in hospital is less than 5 so there were 40 such patients. When the

insurance cost covered was more than 50% by the insurance company the length of stay in hospital for more than 10 days there were 190 such patients. So the total number of patients on which this survey was conducted was 660.

So null hypothesis is length of stay and type of insurance are independent as we are saying it is a chi-square test of independence right so every time you have to use independent over here in null hypothesis statement. So expected frequency row total * column total/n.

(Refer Slide Time: 29:56)



Let us look at this. So this is how you are supposed to use formula for calculating expected for calculating chi-square statistics. So in this question chi-square statistics is 24.315 and the critical value from table when alpha=0.01 is 13.27 so should we reject or should we not reject this hypothesis tell me. This is your critical value right which is 13.277 and observed value is this.

So we will reject the null hypothesis is not it we will reject the null hypothesis it means what length of stay and type of insurance are related with each other they are dependent on each other right. So let us solve this question using Minitab. We will solve this question and then we will wind up today session. So just first delete the output of previous question. So let us look at length of stay less than 25.

Yeah the cost covered by insurance company less than 25% to 50% and more than 50. Here on the stay is less than 5 days to 10 days and more than 10 days right. So you have got 40, 75, 65 then 30, 45, 75, 40, 190 right. So just go to stat table chi-square test for association you

just select all these columns and go to statistics and of course okay. So this is what is P calculated 24.316 yes just see 24.315 is this and are we rejecting null hypothesis?

Yes, we are rejecting null hypothesis. Let us look at P value and confirm whether we are really null hypothesis or not. So let us look at P value yeah this P value is 0 right. So alpha is 0.01 P is 0. So P is less than alpha so we will reject null hypothesis right. So this is how you can solve the questions using Minitab on chi-square test of independence. So with this let me complete today session. Thank you very much.