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

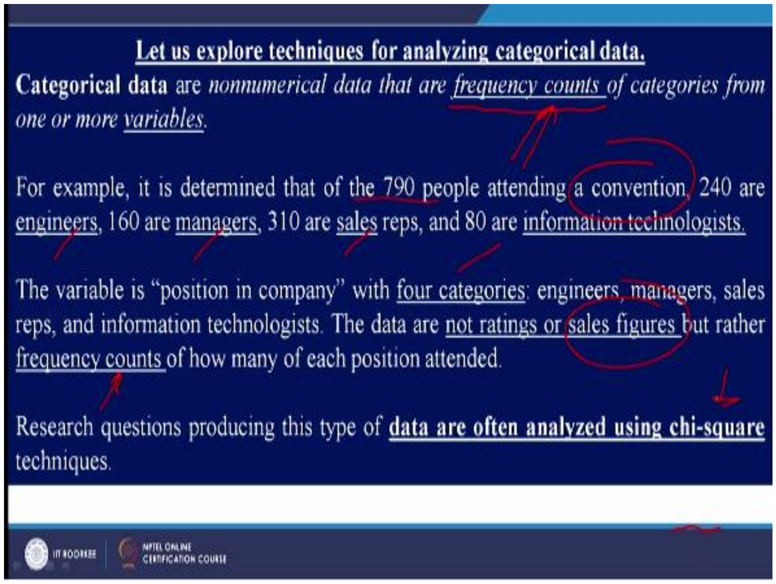
Lecture - 53
Chi-square Test: Goodness of Fit – I

Hello friends, I welcome you all in this session. As you are aware in previous session we were discussing about analysis of variance and we have seen several types of ANOVA, we have seen one-way ANOVA, we have seen two-way ANOVA. So we can conclude that there are three types of design of experiment. The first one was completely Randomized design the other one was Randomized block design and the third one was Factorial design.

Let us start a new topic which is on Chi-square test. Now we have seen in initial classes that there are four types of scales ordinal, nominal, interval and ratio scales. And we have seen several methods of analyzing data whenever we collect data on one particular type of scale. Now let us look at one more technique it is called Chi-square technique for analyzing data which are non-metric in nature or they are categorical data.

For example, we have said that categorical data is like group 1, group 2, married, unmarried, bachelor and so on, right. So you can have, let us say male, female; you can have let us say heavy coffee liquor, medium coffee liquor, right, so all these are nothing but different categorical data or non-metric data, right. So when you have got non-metric data, non-numerical data then you need to analyze the data, right.

(Refer Slide Time: 02:24)

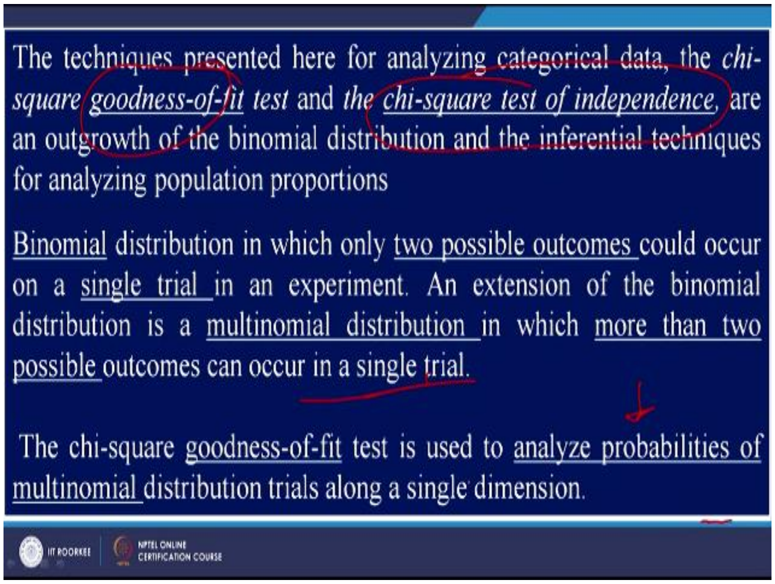


So let us say when you got non-metric data like frequency counts of one or more variables then we need to have a technique for analyzing such kind of data, right. So the example is this let us say, it is determine that of 790 people attending a convention, 240 are engineers, 160 managers, 310 sales representatives, 80 are information technologists. So we will say that there are four different categories of a particular variable which is let us say they are attending a convention.

So these are data which are not on interval scale or which are not on ratio scale or not on rating scale, right. They are not, let us say sales figures they are not income data or age of a person or weight of a person, height of the person and so on, right. So these are frequency counts. And to analyze such data there is a technique called Chi-square technique. And we will see how this technique can be applied in real life or in several other business applications.

So the Chi-square techniques can be broadly classified as Chi-square goodness-of-fit and chi-square test of independence. So basically we will be looking at couple of examples.

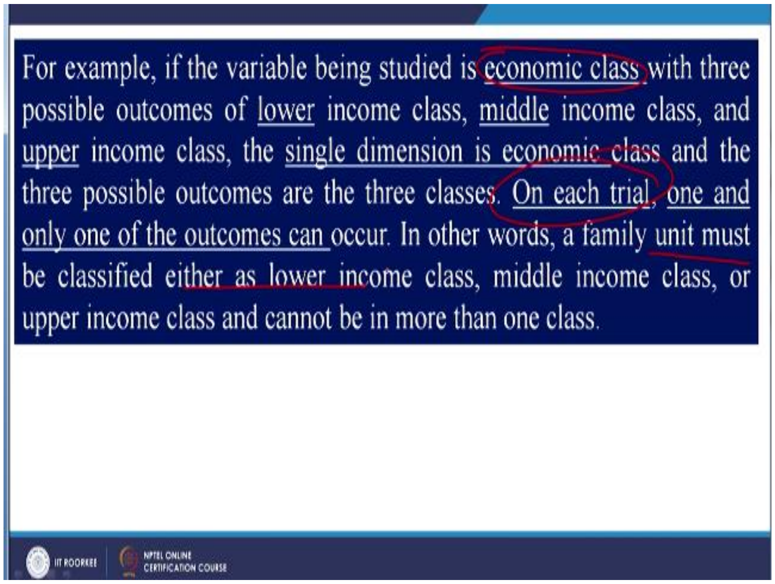
(Refer Slide Time: 04:06)



On goodness-of-fit and couple of examples on chi-square test of independence, right. So we know that in binomial distribution we have got only two possible outcomes of a single trial for example, success, failure or yes or no so on, right. So the extension of binomial distribution is multinomial distribution, right where you will have more than two possible outcomes of a single trial, right.

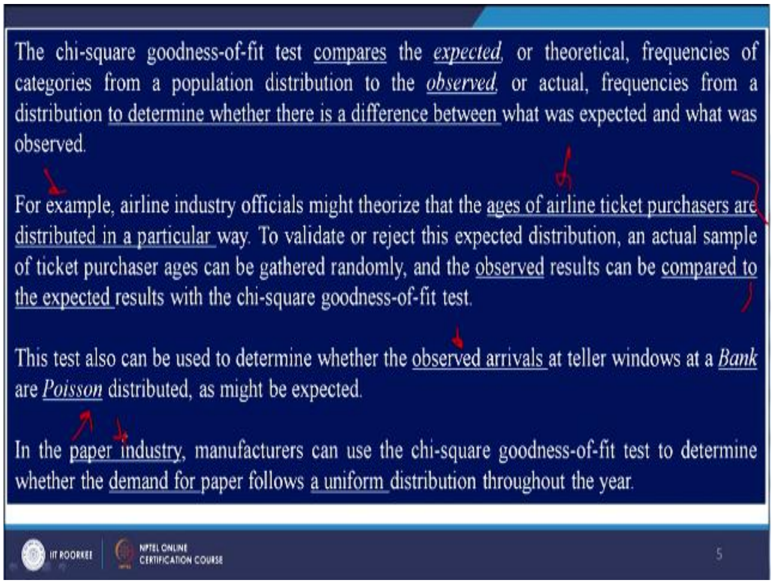
So chi-square goodness-of-fit test is use to analyze probabilities of multinomial distribution trials along a single dimension. So here we will have a trial and the outcomes would be more than two, right.

(Refer Slide Time: 04:58)



So let us say the variable is economic class and we can say that some people are some people have got lower income class, middle income class and high income or upper income class, right. So we will say that there are three different classes, right. So whenever we go for a trial we will have either one of these, right. You cannot have any other class, because initially you have divided three classes, right. So we can say that a family unit must be classified either as lower income, middle income or upper income and cannot be more than one class, right.

(Refer Slide Time: 05:40)



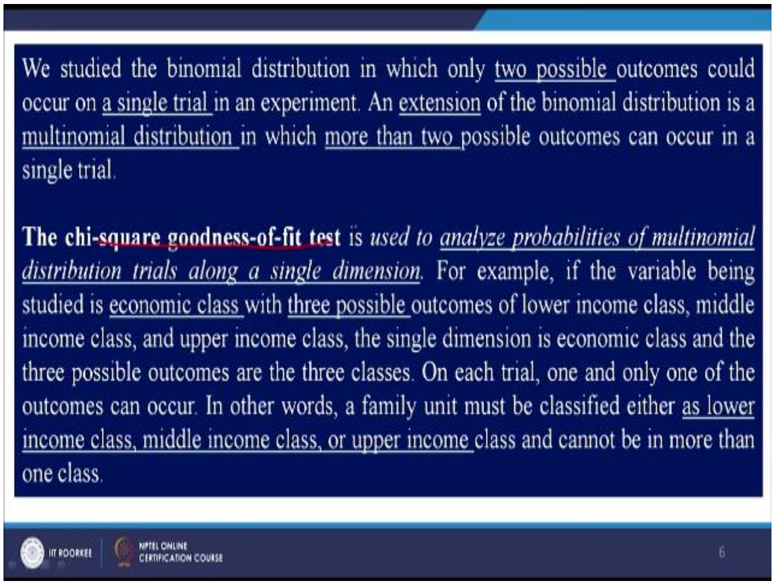
So chi-square goodness-of-fit is basically compares the expected frequencies or expected values with observed values. How expected frequencies match with the observed frequency? So we just check whether there is a significance difference between expected frequency and observed frequency. So ideally if there, if they are same then the goodness-of-fit value should be 1, right. So let us look at this one example. Let us say the managers are the top executives of the airline industry thing that the, the ticket purchase are distributed in a particular way.

Whether this distribution can be a normally distributed or let us say exponential distribution or some other distribution, right. So they have, the managers expect that the ticket distribution is following certain distribution right, certain probability distribution. So they want to you know, collect data and they want to match whether the expected distribution and observed distribution are one and the same or not.

So to solve a question like this we will use chi-square goodness-of-fit test, right or let us say, it can also be used to determine whether the observed arrivals at teller windows at a bank are Poisson distribution, right. So we assume that the arrival at a particular teller of a bank or at a particular tollbooth on a highway is Poisson distributed, right. And then we will collect data and we will check whether it is really Poisson distribution or not, right, or whether the arrival is really Poisson distributed or not or in paper industry the demand is easily uniformly distributed throughout the year.

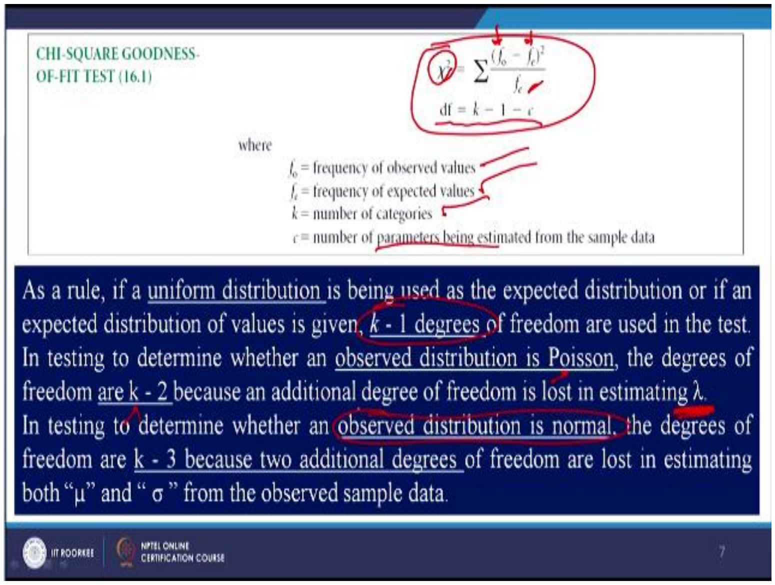
So such expected, expectations are such theories, such theories really have got support from actual data, right that is what we want to prove, right. That is what I said, it compares expected values or you can call them theoretically frequency with observed frequency or actual frequency, right.

(Refer Slide Time: 08:36)



This is how you can have chi-square goodness-of-fit. It analyze as we have already said analyze probabilities of multinomial distribution, right.

(Refer Slide Time: 08:57)

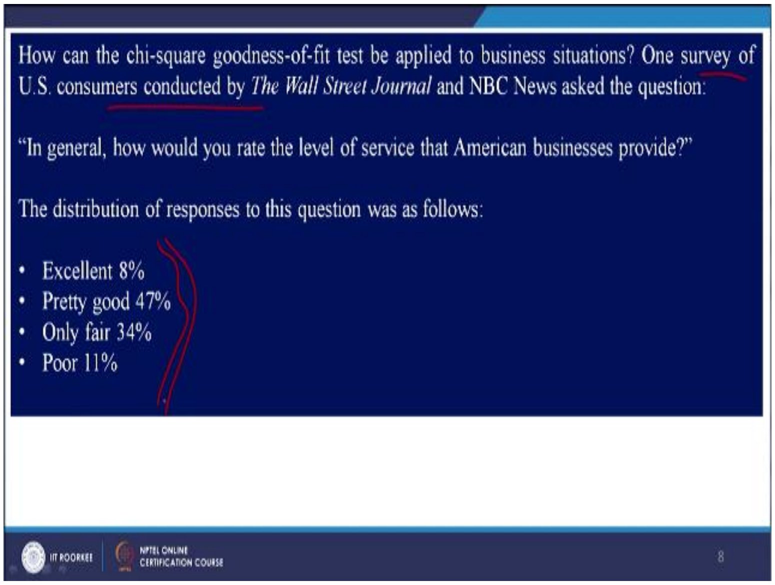


So let us look at how it works. So chi-square goodness-of-fit this is how the formula of that chi-square goodness-of-fit is, so it is chi-square = summation of observed frequencies – expected frequencies and square of it right, divided by expected frequency at this much degrees of freedom, right. So fo observed frequency, fe expected frequency, k is number of categories and c is number of parameters, right. So as a general rule if the distribution is, is uniform then we will have the k-1 degrees of freedom are used in the test.

But if the distribution is Poisson one right, if you think that the arrival of customers at a bank is distributed in Poisson manner then the degrees of freedom would be k-2 because there would be loss of 1 additional degree in calculation of mean value, right, this lambda, right. This is mean arrival rate in Poisson distribution, right. So and if the distribution is let us say a normal distribution, right.

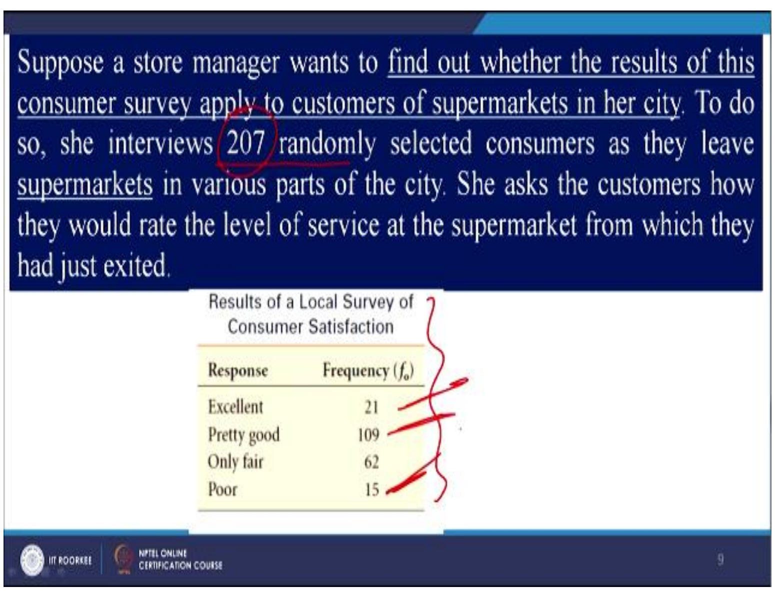
If you think that the expected distribution is normal distributed then degrees of freedom are k-3 because two additional degrees of freedom are lost, one in calculation of mean and other one in calculation of standard deviation, right, so mean and standard deviation. So it is very simple. If it is a Poisson; if expected frequency is like Poisson distributed then one degree of freedom is lost; if it is; if expected distribution is normal distribution then two degrees of freedom is lost, right. So let us look at an example.

(Refer Slide Time: 11:10)



In a survey of U.S. consumers conducted by Wall Street Journal and NBC News asked the question to the customers and the question was, in general how would you rate the level of service that American businesses provided? And there were four options in this question. And, this was the finding of the question. So 8% said Excellent, 47% said Pretty good, 34% said Only fair and 11% said Poor, right. So this was a survey conducted earlier by let us say these two agencies.

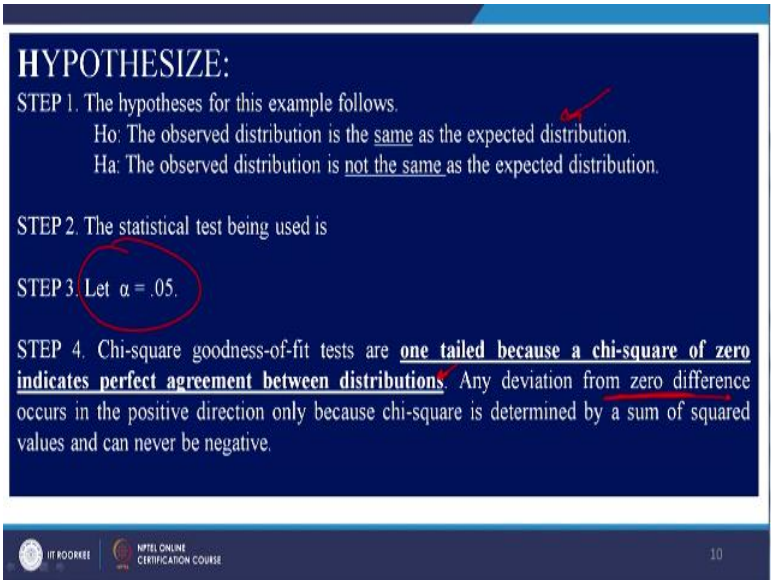
(Refer Slide Time: 11:56)



Now, the store manager of a particular city wants to know whether the behavior of the customer is similar to the behavior of the customer at National level, right. Okay. So this was the national level data. So what she has done is, to do so she interviews 207 randomly selected consumers as they leave supermarkets in various parts of the city and she asks the same question, right and this the response she has got, right. So out of 207; these are the frequencies, right. So 21 said Excellent, 109 said Pretty good, and 15 said poor, right.

So we want to know whether the observed frequencies are same as expected frequencies. That is the question.

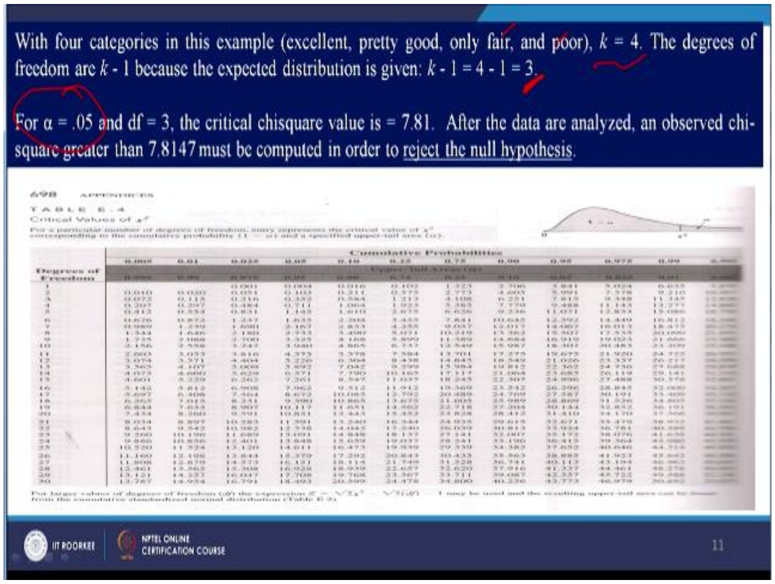
(Refer Slide Time: 13:00)



So we will have hypothesis testing. So we will say that our null hypothesis is that the observed distribution is same as expected distribution and alternative is they are not same, right. The expected distribution is different and observed distribution is different. So we will use test statistical. At alpha=0.05, right. So it is; keep in mind that chi-square test is always one tail test, because a chi-square of 0 indicate perfect agreement between distribution.

So keep in mind that if the expected distribution and observed distribution; if the difference is 0 then we will say that they are one and the same, right. So any deviation from 0 in a positive direction only because chi-square is determined by sum of squared values and can never be negative, because we are saying it is a chi-square test, right. The moment we say square, so it cannot be a negative value.

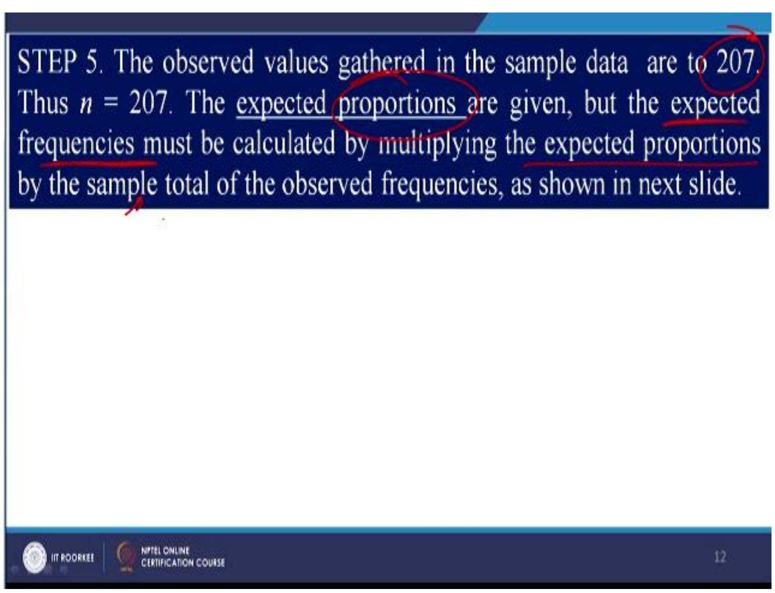
(Refer Slide Time: 14:20)



So in this example which we have taken up there are four categories, so Excellent, Pretty good, Only fair and Poor, right. So k=4. So the degrees of freedom are k-1 as we have said, so at 3 degrees of freedom at alpha=0.05 we have to see the chi-square value, right the table value of chi-square, right. so let us look at 3 degrees of freedom here, right and at 0.05 alpha value. So this comes out to be 7.815, is not it. so this is the table value, right.

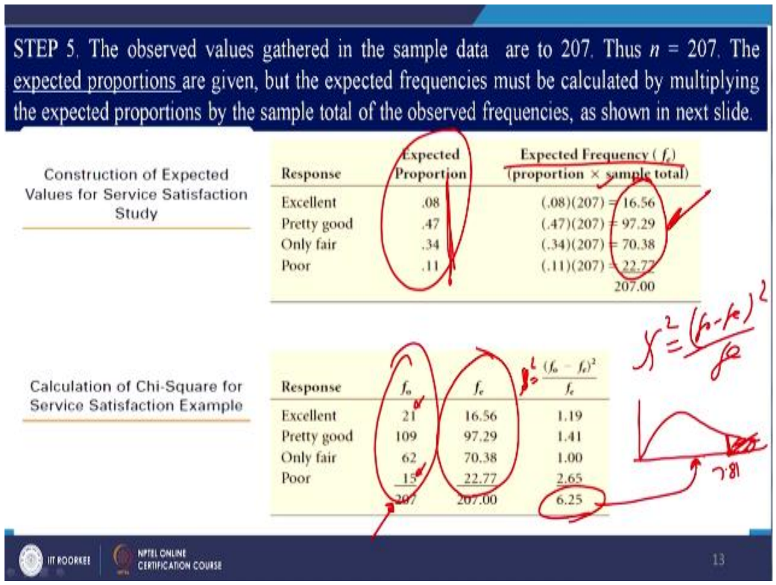
Now, you need to calculate chi-square value and then you need to compare these two whether, means to decide whether you should reject or not reject null hypothesis. So the next one is, we need to have the expected proportion.

(Refer Slide Time: 15:27)



So as we know that the manager of the supermarket collected data from 207 customers, so we need to calculate now expected frequencies, right. In this question we have been given the expected proportions, right not the frequencies, right. So first we convert, we should convert expected proportions into expected frequencies, by multiplying the expected proportion by the total sample size.

(Refer Slide Time: 16:03)

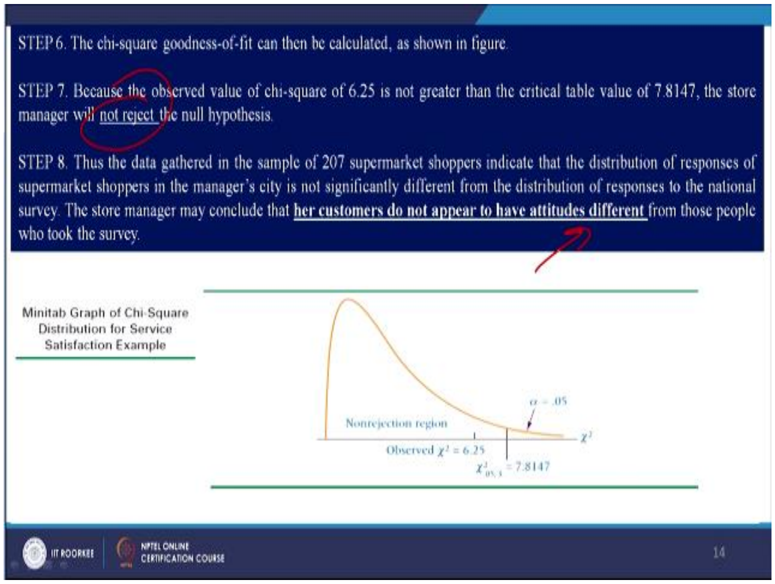


So we know that these are expected proportions given in the question initially, right. and we know that the two find out expected frequency you need to multiply these values by 207, right each of these values will be 207. So these are now expected values, right. These are expected values are, you can all it expected frequencies, f_s. So now, these are expected frequencies and these are observed frequencies, right is not it.

That is what that manager found out that out of 207, 27 said excellent and 15 said poor, right. So is there any significance difference between these two fo and fe, right. So calculate chi-square value, this is how chi-square value would be like this. Okay let me redraw it. Okay. So chi-square is this, right $\frac{(f_o - f_e)^2}{f_e}$ right. So the chi-square value for this is 6.25. And the table value was it was 7.81, right. So how would you distribution?

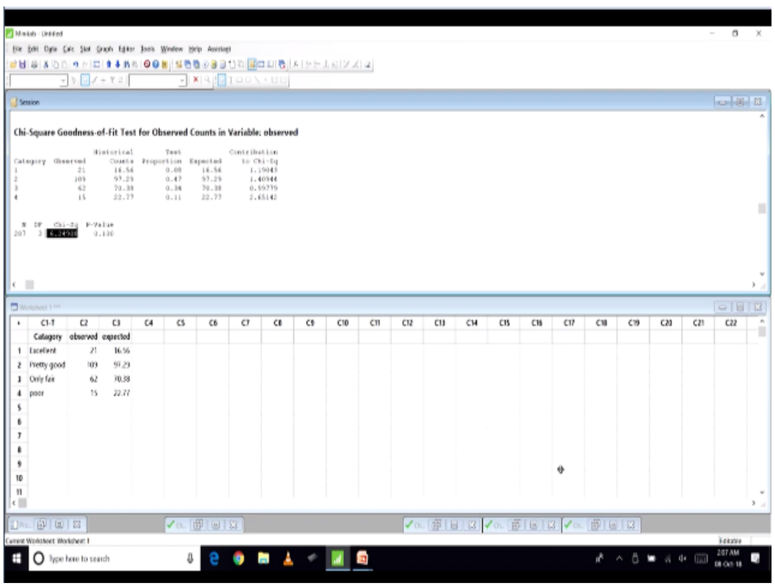
So this is 7.81, right your critical value, and we calculated and this is your rejection region, right. This is non-rejection region. So 6.25 is somewhere here, right. So you will not reject the null hypothesis. It means what, it means that the national level or the customers national level and the customers of the city think in similar manner, right. There is no significant difference, right.

(Refer Slide Time: 18:19)



So in this case we will not reject the null hypothesis and we will say that the customers do not appear to have attitudes different from the national level, right. So this is your rejection region and this is your non-rejection region, right. Let us look at the sample example using Minitab, right. So we will have this data, right and we will use Minitab software, right.

(Refer Slide Time: 18:47)



So first of all here you can write the categories, right. So expected value and observed value, right. So excellent, pretty good, only fair and poor right, so we know expected and observed frequencies so 21, 109, 62, 15, 16.56, 97.29, 70.38 and 22.77, right. So to solve a question like this using chi-square test go to Stat, go to Tables then Chi-square goodness-of-fit, right one variable, right. Again, chi-square goodness-of-fit, right.

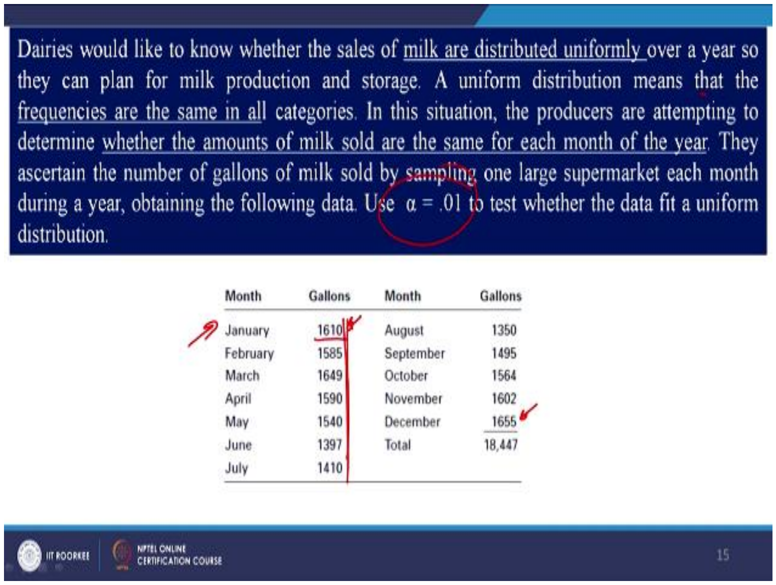
So you have got expected and observed frequencies here, so this observed count is C3 okay and category name is this category right. You do not do anything over here, right. Proportions specified by historical counts, right. So this is input column is expected, right. So you have got observed over here and this is expected. Let us look at next point. So let us see what are different graphs available, so we can; otherwise we can directly click OK. And, OK.

In fact, these were the observed and the other; yeah in fact that is what you need to keep in mind that there are some observed and some of them are expected, right. So these are observed and these are expected, right. Chi-square goodness-of-fit, Observed, Expected, right. OK. So let us look at P value. P value is 0.10. So let us go to our question. So P value is 0.10 and the alpha value was 0.05, right. So is P value less than alpha?

No. So do not reject null hypothesis, right. So this is the conclusion from output of chi-square table, right. Always look at P value. So P is greater than alpha so we do not reject null hypothesis. And we will say that the attitudes of customers of a particular city and the attitude of customers at national level are same, right. So the other values you can see over here.

In fact, you can look at the total observation 107 at 3 degrees of freedom and this what we calculated, right 6.24. Just check. 6.24 was the calculated chi-square value, right 6.24 not 25 one and the same thing, right. So this is how you can solve a question using Minitab on Chi-square. So let us look at one more example.

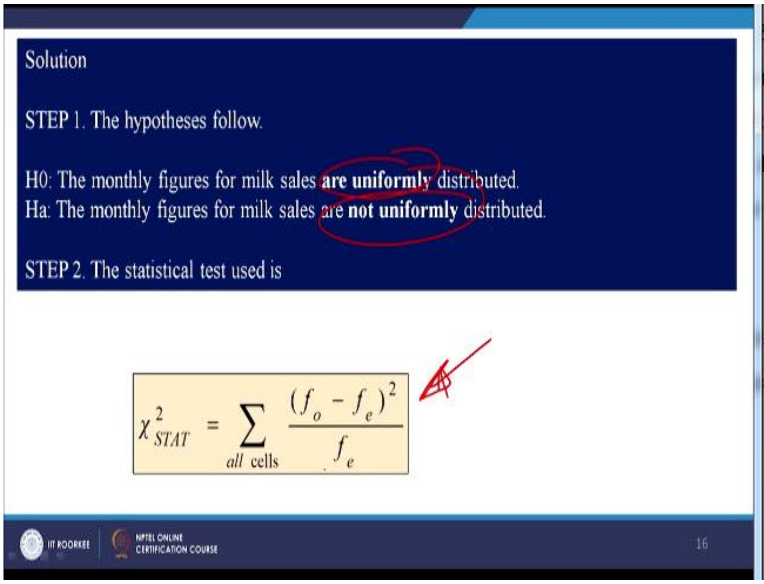
(Refer Slide Time: 23:37)



Let us say Dairies would like to know whether the sales of milk are distributed uniformly over a period of 12 months, right or over a period of one year. A uniform distribution means that the frequencies are same in all categories, right. So in this situation, the producers are attempting to determine whether the amounts of milk sold are same for each month of the year.

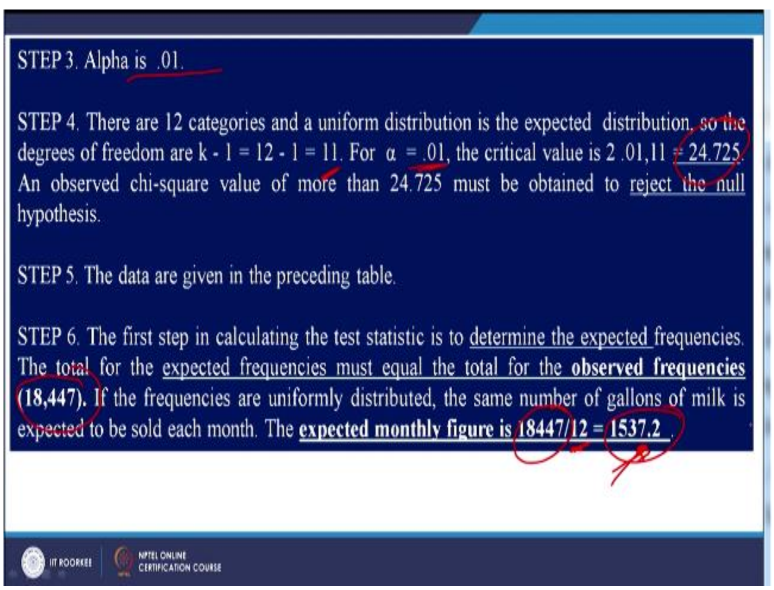
So they have collected data from dairies of each of each of 12 months of the year and these were the data they have collected, right. So in the month of January this is the consumption of milk 1610 gallons. In the month of December, the consumption is 1655 gallons. So the milk dairies people they think that the distribution or the consumption of milk is uniform throughout the year and this is to be tested at 0.1 significance level, so how to proceed this question?

(Refer Slide Time: 24:49)



So you will have null and alternative hypothesis so you will say that the distribution is uniform, right and it is not uniform is alternative hypothesis, this how you can use chi-square statistics, you can calculate like this.

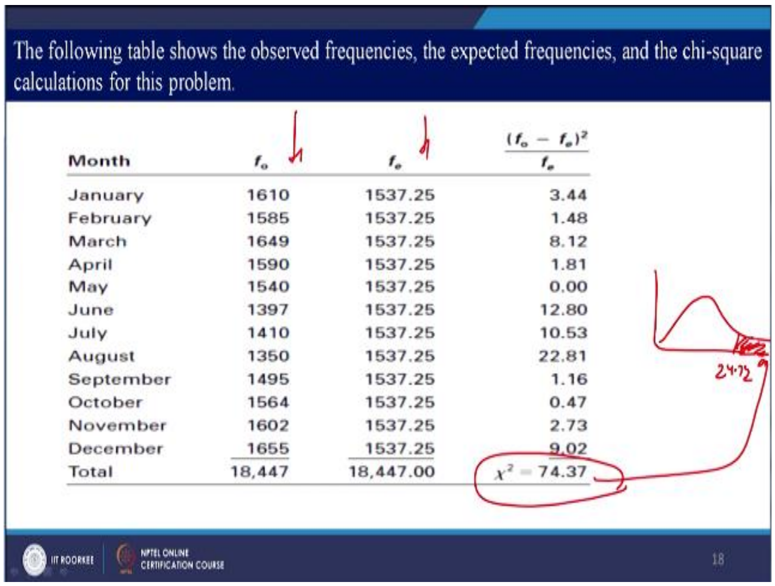
(Refer Slide Time: 25:07)



You know alpha here is 0.01. So there are 12 categories. So at 11 degrees of freedom when alpha=0.01 the table value is 24.72, right. And if you want to reject null hypothesis the calculated value should be more than this, right. So let us look at the observed frequencies and expected frequencies. So in the question itself we have been given these are observed frequency, right. and the expected frequencies would be calculated by totaling the demands of all the 12 months divided by 12, right 12 months are there, right.

So this is expected frequency of a particular month. And this will remain same for all other 11 months. So the total expected frequencies must be equal to the total observed frequencies, right. So this is observed frequencies, 18447. How did you get this value? This is here, summation of all these values. So 18447 is the observed frequencies, right summation of all the fo and the fe would be 18447/12, right. So that would be 1537.2, right.

(Refer Slide Time: 26:46)



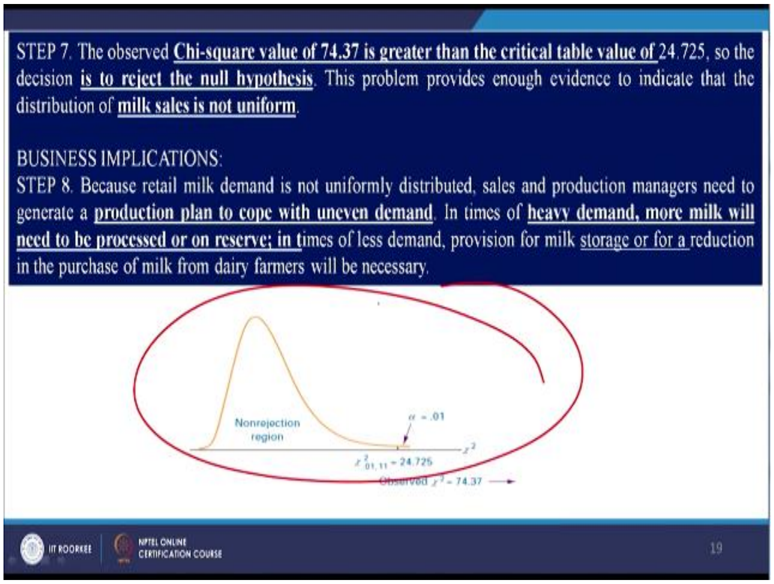
The following table shows the observed frequencies, the expected frequencies, and the chi-square calculations for this problem.

Month	f_o	f_e	$\frac{(f_o - f_e)^2}{f_e}$
January	1610	1537.25	3.44
February	1585	1537.25	1.48
March	1649	1537.25	8.12
April	1590	1537.25	1.81
May	1540	1537.25	0.00
June	1397	1537.25	12.80
July	1410	1537.25	10.53
August	1350	1537.25	22.81
September	1495	1537.25	1.16
October	1564	1537.25	0.47
November	1602	1537.25	2.73
December	1655	1537.25	9.02
Total	18,447	18,447.00	$\chi^2 = 74.37$

So you have got all fo over here, all fe over here and this is your calculated chi-square value, right. So this is what you have calculated and critical value is what, it is what we have found out from table, right. It is 24.72 and the calculated is somewhere here, right in rejection region. So we will reject the null hypothesis. What was null hypothesis? That the distribution is uniform but we are rejecting it.

It means now the distribution of milk is not uniform across the year, right. It is different. It means in some month it is more than average; in some months it is less than average, right.

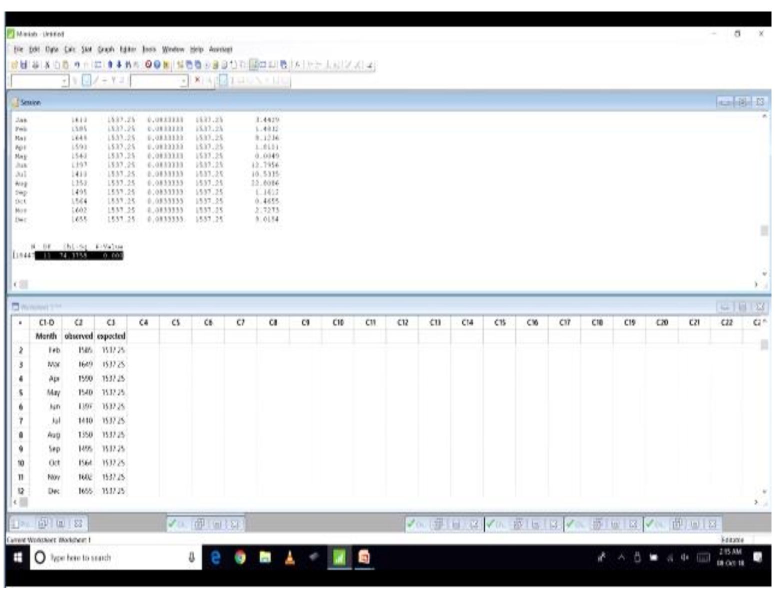
(Refer Slide Time: 27:39)



So this is how your distribution would look like. And would be the business implication of this? We know that the distribution is not equal in some months it is more. So when in some month it is more distribution it means you need to arrange milk from somewhere, right. And in some month’s demand is less. So when the demand is less you need to have some storage facility, right.

So this is how management can take a decision let it to either how they store the milk or where to arrange the milk when the demand is more than average, okay. So let us look at this question and solve using Minitab.

(Refer Slide Time: 28:32)



So we will look at the; this particular question. So you have got different months Jan, Feb, March, April, May, June. So you can write all these months over here from January to December, okay Sep will be; okay October, Nov-November and Dec-December, right. You just enter your observed values over here, 1610. So you need to enter all these values so far as observed frequencies are concerned.

And we know that the expected frequencies is the summation of all the; is the summation of all the observed frequencies and divided by 12. So we know that the expected frequencies is same, it will remain same in all these months, right. So you can just copy this value and paste in remaining cells, right. Now you just go to Stat, Table, Chi-square goodness-of-fit, Observed, Category is month, input column is expected, right and click OK.

So look at P value, P value is less than alpha, so we will reject the null hypothesis that is what we concluded earlier and calculated P value is 74.37, just check whether we are getting the same value, yes 74.37. And N is 18447, right. So in this way you can calculate questions on chi-square test as far as goodness-of-fit is concerned, right. With this let me stop over here. And we will have few more questions on chi-square test in next session. Thank you.