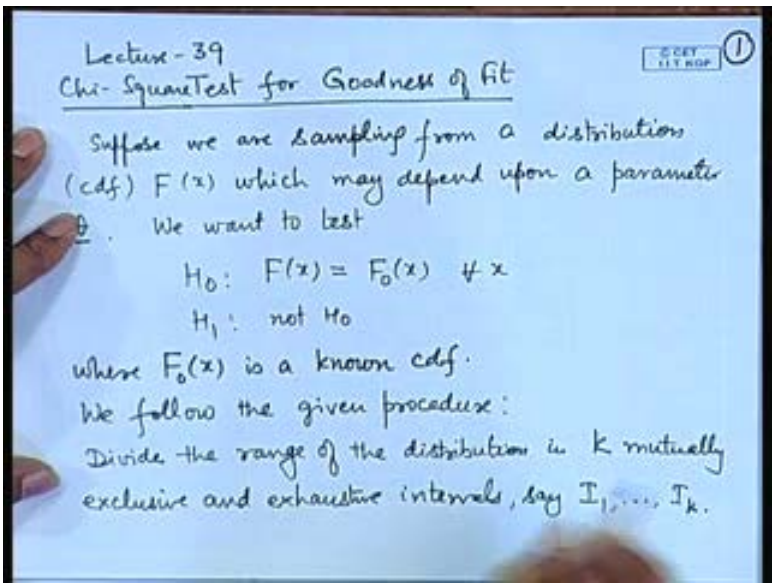
Probability and Statistics
Prof. Dr. Somesh Kumar
Department of Mathematics
Indian Institute of Technology, Kharagpur

Module No. #01
Lecture No. #39
Testing of Hypothesis-VII

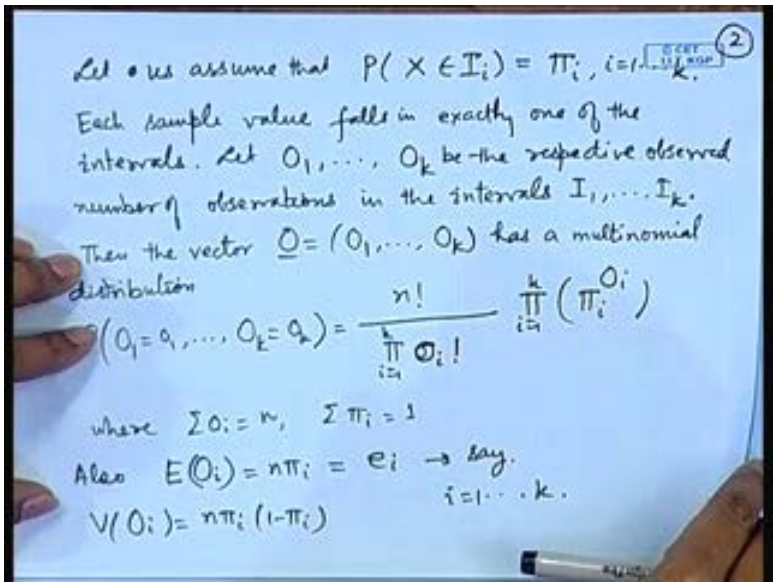
In the situations so far, for testing of hypothesis problems, we have considered that the random sample comes from a certain population. So, we have assumed the form of population to be say, normal or say exponential. And then, we want to test about the parameters of that population. For example, we want to test about the mean of a normal distribution, we want to test about the variance of a normal distribution or we want to test about the scale parameter of an exponential distribution. So, here the situation is that, we assume that the form of the distribution is known to us. Only the parameters of the population are not known. However, there are other situations where we have a data and we want to know that, from which type of population that data has come. And so, that means, we may be liking to estimate the population, that means, the distribution or we want to test about the distribution. So, here we will talk about the testing, that a particular distribution is there, say capital F is equal to sum F naught. So, for this situation an approximate test has been proposed which is called chi square test for goodness of fit.

(Refer Slide Time: 01:47)



So, let us have the situation of the form, suppose we are sampling from a distribution. So, let me write distribution function c d f say F x. Of course, there may be a situation where it depends upon certain parameters. So, we may write that thing, which may depend upon a parameter say theta. So, this theta could be having several components also. So, we want to test, say H naught F x is equal to F naught x, for all x against H 1 not H naught. That means, this is not true for some points at least. Where F naught is a known c d f, that means, we want to test whether the data which has been collected comes from a given distribution F naught x. So, in the chi square test for goodness of fit the procedure is as follows: We follow the given procedure. So, divide the range of the distribution in k mutually exclusive and exhaustive intervals. Let us call them intervals as I 1, I 2, I k.

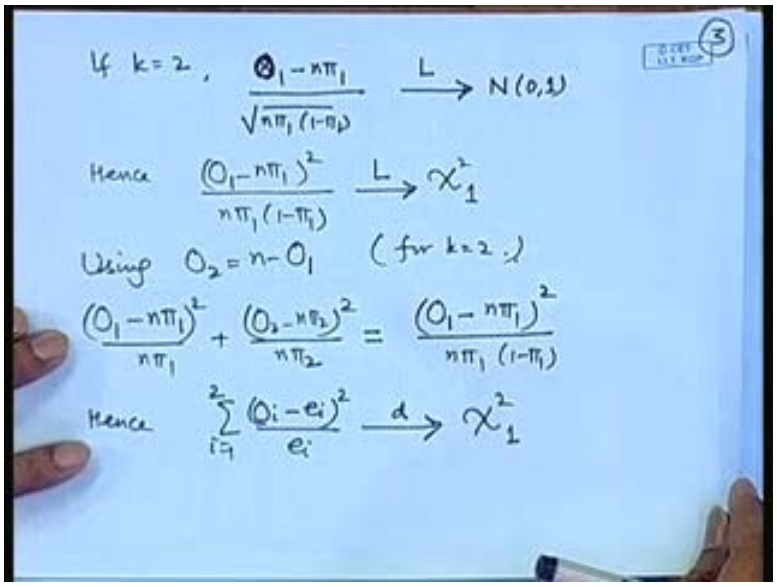
(Refer Slide Time: 04:31)



So, now, each value will fall exactly in one of the intervals because, they are mutually exclusive and exhaustive. Let us also assume that, let us assume that probability of x being in an interval I is pi i for i is equal to 1 to k. So, now each sample value falls in exactly one of the intervals. So, let us define the observed frequencies. Let O 1, O 2, O k be the respective observed number of observations in the intervals I 1, I 2, I k. So, what we are observing? We are observing x 1, x 2, x n. Now, you see some of these x i's will belong to interval I 1, some of the x i's will belong to interval I2 and so on. So, we make this break up.

Let O 1 denote the number of x i's in the interval I 1, let O 2 the number of x i's in I 2, let O k denote the number of x i's in I k. Then, O 1, O 2, O k these are called observed frequencies of the data. So, now, you see here. You have k categories unlike binomial distributions, where you have 2 categories, here you have k categories. So, if I find out the distribution of O 1, O 2, O k, this will become a multinomial distribution. So, we will write it in this form, then the vector O, that is O 1, O 2, O k has a multinomial distribution. n factorial divided by product of O i factorial, then product of pi i to the power O I, i is equal to 1 to k where sigma of O i is n sigma of pi i is 1. That is pi i is the probability of the ith interval. Therefore, probability of O i is equal to small o i, that will be given by this multinomial function. Also, we will have from the properties of the multinomial distribution, expectation of O I, this become equal to n pi I, we call it e i. And variance of O I, that will become equal to n pi i into 1 minus pi I, for i is equal to 1 to k.

(Refer Slide Time: 08:22)

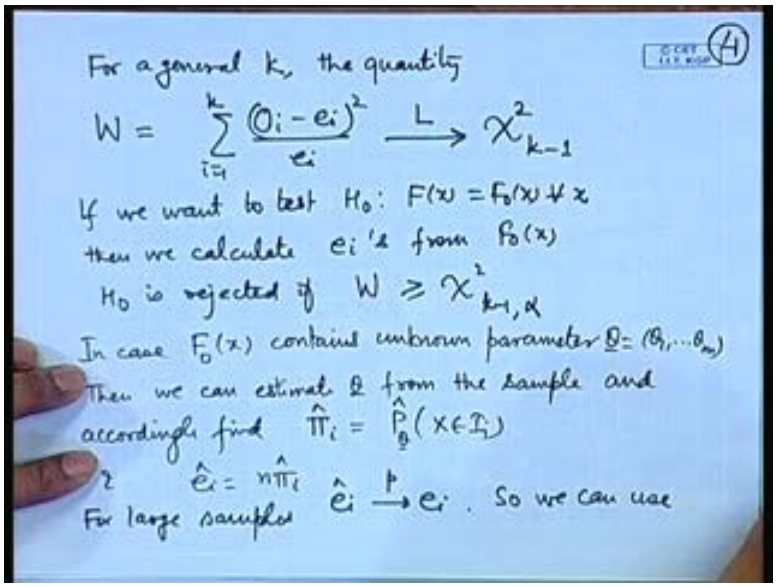


Let us take the case of 2 categories. Then, let us see how the test can be conducted. So, if k is equal to 2, then if I look at x 1 minus n pi 1, in place of this I will write O 1. So, O 1 minus n pi 1 divided by root n pi 1 into 1 minus pi 1. This is converging in distribution to normal 0 1. This is from the property of the binomial distribution that the distribution of x minus n p divided by root n p q is asymptotically normal. That is the normal approximation to the binomial distribution. So, if we utilize that because for k equal to 2,

this multinomial has become binomial. Therefore, the distribution of O 1 is binomial n pi 1. So, O 1 minus n pi 1 divided by square root n pi 1 into 1 minus pi 1 is asymptotically normal distribution. So, if I take the square of this, then this is asymptotically chi square distribution on 1 degree of freedom, because square of a normal distribution, standard normal variable is a chi square variable.

Now, O 2 is n minus O 1. This is for the case k equal to 2. So, you can easily see that, if I write down O 1 minus n pi 1 square by n pi 1 plus O 2 minus n pi 2 square by n pi 2, then, so, here you substitute O 2 is equal to n minus O 1 and pi 2 is equal to 1 minus pi 1. Then, after simplification because, this will become 1 minus pi 1. We can take LCM and adjust the terms. This will give simply O 1 minus n pi 1 square divided by n pi 1 into 1 minus pi 1. So, what we are observing, that sigma O i and this time I call e I, square by e i, i is equal to 1 to, this is having asymptotically chi square distribution on 1 degree of freedom.

(Refer Slide Time: 11:03)

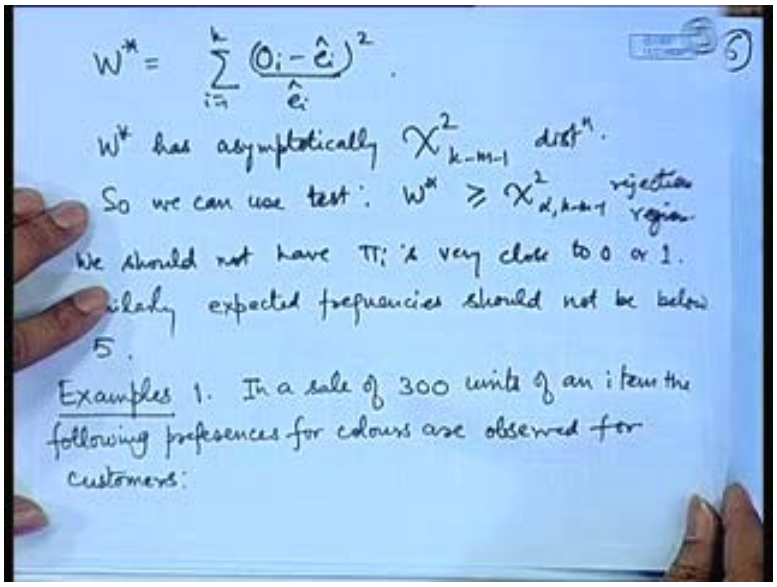


So, if we generalize this thing, in place of 2, if I write for a general k, the quantity, let me call it W. That is equal to sigma O i minus e i square by e i, i is equal to 1 to k. This is having asymptotically chi square distribution on k minus 1 degrees of freedom. I am using L and d as for the asymptotic distribution here.

So, if we want to test the hypothesis, if we want to test H naught F x is equal to F naught x, then we calculate e i's from F naught x. That is, under the distribution F naught what is the probability of the ith interval. So, that is pi i and if I multiply by n, i will get e i. But, of course, this is calculated from the from the known distribution, this one here. So, you can, then consider the difference between the observed frequency and the expected frequency, squared and divided by the expected frequency. So, you can see that, if this hypothesis is true, then the differences between O i's and e i's must be small. And therefore, this term should be rather small. Therefore, by comparing with the tabulated value of a chi square distribution on k minus 1 degrees of freedom, we can test whether H naught can be rejected or cannot be rejected. If it is not true, then these differences will tend to be large. So, the value of W will be large. So, the test H naught is rejected if W is greater than or equal to chi square k minus 1 alpha.

Now, there may be a situation where F naught may not be completely known. That means, it may include certain parameter. If this includes certain parameter, then from the data we can estimate that parameter also. And then in place of e I, we can say we are getting e i heads. And we can substitute there, in case F naught x contains unknown parameter, say theta is equal to theta 1, theta 2, theta n. Then, we can estimate theta from the sample. And accordingly, find pi i head is equal to probability x belonging to I head. That means, the estimate of this. And e i head is equal to n times pi i head. Then, for large samples, e i heads will converge to e i in probability or with probability 1. So, we can use, say W star that is equal to sigma O i minus e i head square by e i head, i is equal to 1 to k. W star has asymptotic chi square k minus m minus 1 distribution.

(Refer Slide Time: 15:13)



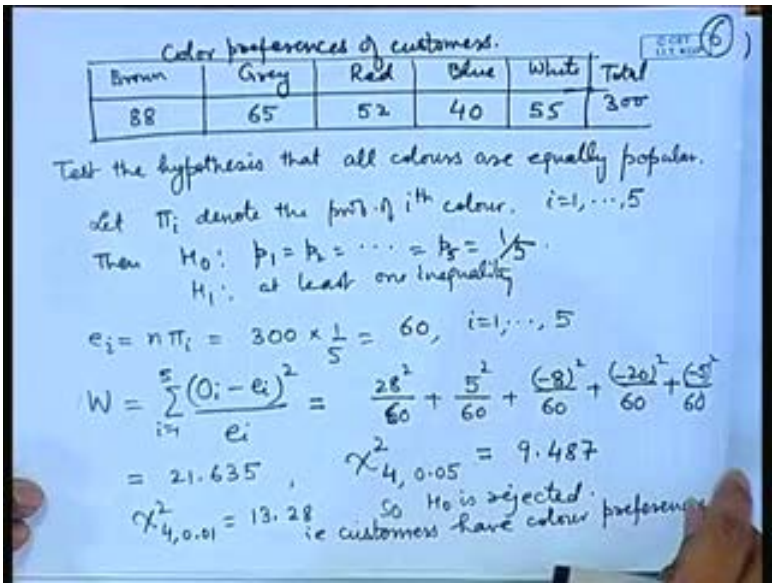
So, once again, we can use test as that W star greater than or equal to chi square alpha k minus m minus 1. This is the rejection region. However, there are certain precautions one should take while using the chi square approximation. As the binomial approximation to the normal distribution is good, when p is moderate; that means, it should not be close to 0 or close to 1. In a similar way, here the sale probabilities are say pi i's. So, if either of the pi i's is extremely small; that means, close to 0 or close to 1, then in that case the expected frequency of that sale will become either too small or too large. If it is too large then for some other sale it may become too small. In that case, this approximation is not good. So, we have the following considerations.

We should not or we can say, we should not have pi i's very close to 0 or 1. Similarly, expected frequencies should not be below 5. So, a practical consideration that has been done, that there may be a case that firstly, we split the intervals without knowing the probabilities but, when we actually calculate the probabilities and find that the expected frequencies are below 5, then, what we do? We can merge some adjacent intervals. So that, the number of interval becomes slightly less but each sale frequency becomes more than 5. So, this is a practical approach that is used here.

Let me explain this test through certain examples. In a sale of say, 300 units of an item, the following preferences for colors are observed for customers. So, it may be like

certain item of the type, say for example, somebody is buying say car or somebody is buying a say two wheeler. So, we look at the color of the car for example.

(Refer Slide Time: 19:21)



So, out of 300 customers, suppose we observe that brown, the colors available are brown grey, red, blue and say white. Out of 300 customers, we find 88 prefer brown, 65 prefer grey, 52 prefer red, 40 prefer blue and 55 prefer white, out of 300. Color preferences of customers. So, we want to test the hypothesis that all colors are equally popular; that means, the customers have equal preference for each of the 5 colors. So, if we want to frame a hypothesis in the form of a test of goodness of fit, what we can do is, the hypothesis is of the form that each sale has probability 0.2.

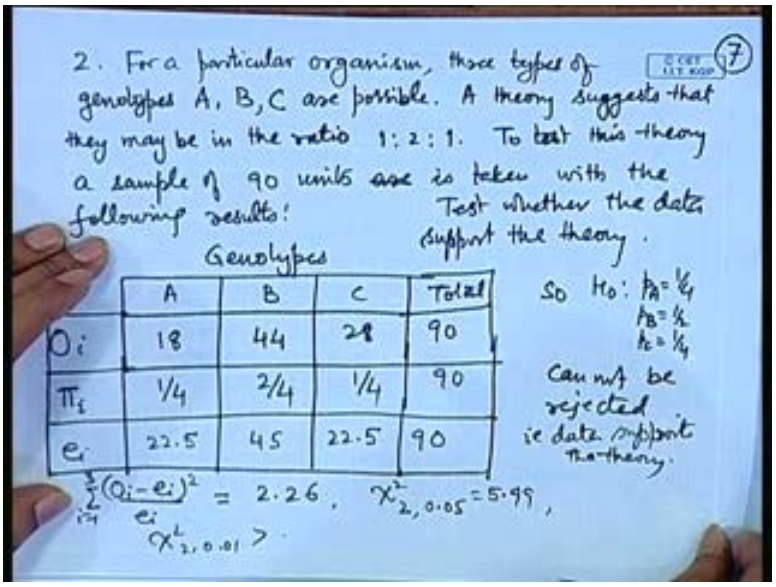
So, let pi i denote the probability of ith color, for i is equal to 1 to 5. There are 5 colors here. Then, we want to test that each of the, is 1 by 5. Now, based on these assumption, so, here basically the sale or interval is actually the type here. So, brown is one type, grey is another type, red is another type, blue is another type, white is another type. So, this is again a multinomial situation and here we are assuming the probabilities to be same in the null hypothesis. H 1 is that, at least 1 inequality. So, on the basis of this, we do the following calculations. We can calculate e i's. So, e i is n pi i. So, here it is 300 into 1 by 5, that is equal to 60. Each category has the same probability and therefore, each category will have the same expected frequency also.

So, on the basis of this if we calculate W, that is sigma O i minus e i square by e i, i is equal to 1 to 5. Then, this is equal to 88 minus 60, that is 28 square by 60 plus 65 minus 60, second sale, that is 5 square by 60, third sale is 52. So, it is minus 8 square by 60 plus 40 minus 60 is minus 20, square by 60 plus 55 minus 60 is minus 5, square by 60. So, the sum can be easily evaluated, it turns out to be 21.635.

Now, there are 5 categories. We also notice that the expected frequency of each sale is more than 5. So, the chi square assumption is valid. Therefore, we look at the value of chi square on 4 degrees of freedom. Suppose, I consider the value at say 0.05, then from the tables of the chi square distribution, one can find this value is 9.487. We may even look at say chi square 4.01, that is 13.28. So, you can see that the calculated value of W, that is 21.635 is bigger than this. So, H naught is rejected; that means, what is the conclusion? The conclusion is that customers have preferences for the colors.

You can see the raw data here. The observed frequency for brown is 88, which is almost more than twice the choice of blue color. If we see the choice of grey, that is much higher. The choice of red, blue and white is below. So, you can see that, in specifically speaking, brown and blue, they cause major discrepancies here. Blue is, say least favorable color and brown is the most favorable color here. In fact, if I have only 3 of these, then they look almost nearby 65, 52 and 55. That is, customers have color preferences. So, basically what we have tested is something like a discrete uniform distribution. And we conclude that the data does not follow a discrete uniform distribution.

(Refer Slide Time: 25:17)

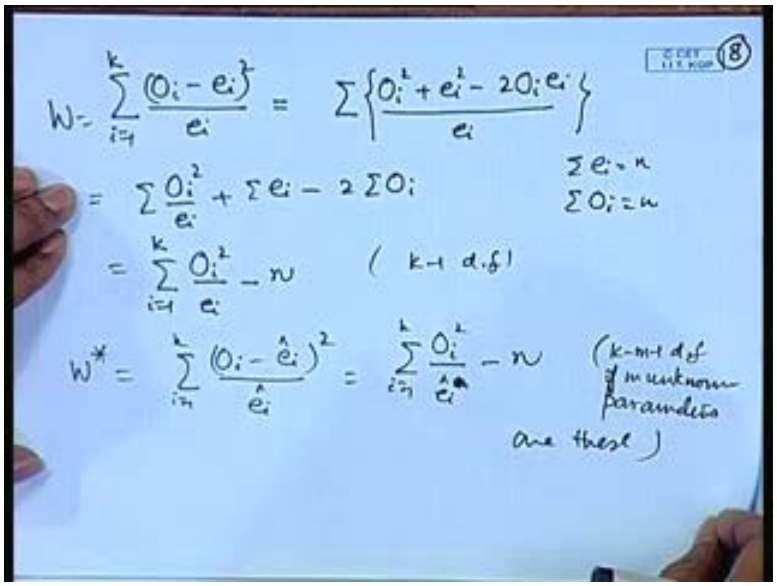


Let us take another example, for a particular organism three types of genotypes A, B and C are possible. A theory suggests that, they may be in the ratio, say 1 is to 2 is to 1. Now, to test this hypothesis, to test this theory, a sample of 90 units are is taken with the following results. So, genotypes A, B, C, it is observed that out of 90 units, 18 had genotype A, 44 had genotype B, and 28 had genotype C. The total is 90. So, now, we want to test whether the data supports the theory. So, for this once again, we have 3 categories. The probabilities of each category, let me call it pi i. So, 1 is to 2 is to 1. So, this probability is 1 by 4, this probability is 2 by 4, that is half and this probability is 1 by 4. So, expected frequency, this is observed frequency. We can actually do the calculations in the form of a table here.

So, here you see, if the probability of the genotype A is 1 by 4, the total number of units is 90. So, the expected frequency for that will be 90 by 4. That is 22.5 here, it will become 45, here it will become 22.5. That is, total is 90. So, based on this, one can carry out the calculations, sigma O i minus e i square by e i, i is equal to 1 to 3. So, for example, here it will become 4.5 square divided by 22.5 plus 1, 44 minus 45. So, 1 square by 45 and 28 by minus 22.5, that is 5.5 square by 22.5. So, one can look at this calculations, this turns out to be 2.26.

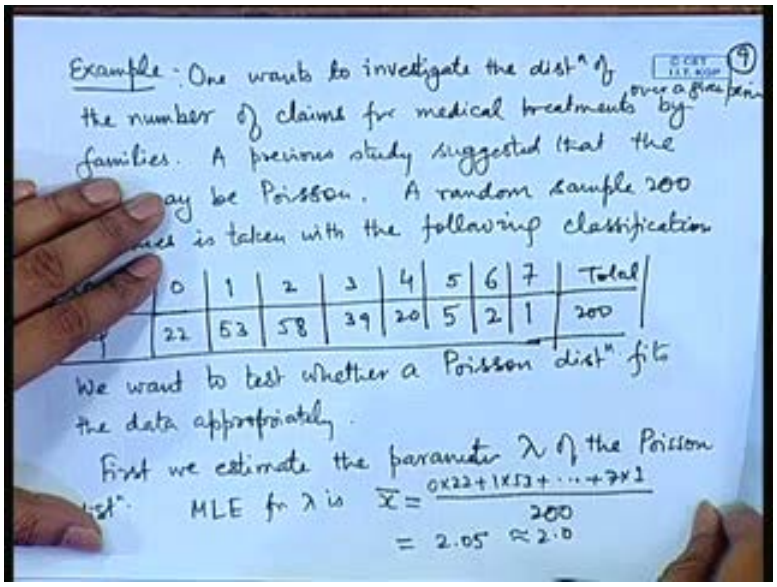
So, one can easily compare with the chi square value on 2 degrees of freedom. Suppose, we look at 0.05, then this is 5.99. And of course, if I look at say chi square 2 at 0.01, this is going to be larger than this. So, we cannot reject H naught; that means, H naught that is p A is equal to 1 by 4, p B is equal to half, p C is equal to 1 by 4, cannot be rejected. That means, the data supports the theory that the genotype are in the proportion 1 is to 2 is to 1.

(Refer Slide Time: 30:06)



Here, one point about this calculation also. This formula which we have given here, sigma O i minus e i square by e I, one can actually have an alternative form for this. We can consider expanding this term. So, it is O i square plus e i square minus twice O i e i divided by e i. That is equal to sigma O i square by e I, now the next term here is e i and then summation. So, summation e i is actually n, minus twice summation O i. So, summation e i is n and summation O i is also n. So, this become simply sigma of O i square by e i minus n, i is equal to 1 to k. So, this is an alternative formula for W star, for W. Similarly, if I am considering W star, that is sigma O i minus e i head square by e i head, then once again, this can also be written as sigma O i square by e i head square minus, sorry, e i head minus n. Here, degrees of freedom are k minus 1 and here the degrees of freedom are k minus m minus 1, if m unknown parameters are there. Many times, this expression is easier to calculate.

(Refer Slide Time: 31:58)

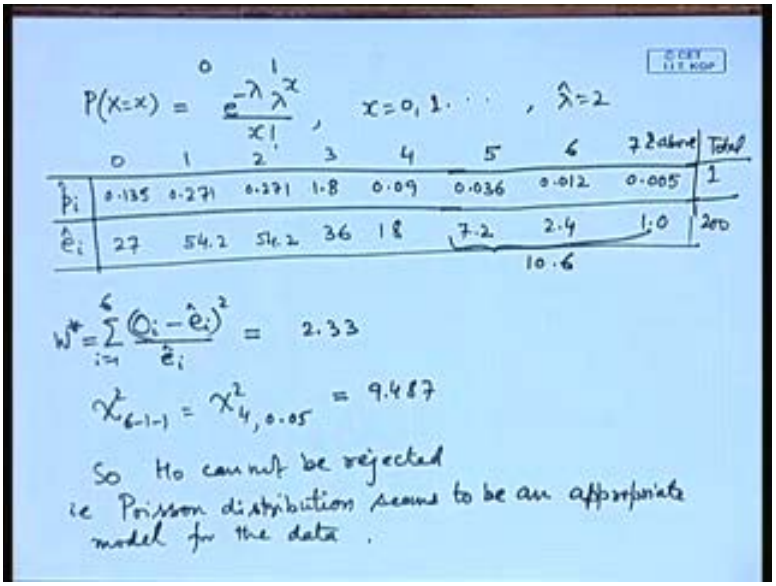


Let us take one case, where the distribution will depend upon certain unknown parameter. So, one wants to investigate the distribution of the number of claims, for medical treatments by families. A previous study suggested that the distribution may be poisson. So, to investigate this, a random sample of 200 families is taken with the following classification. So, that is, for each family how many claims are there? And that is what is the frequency? How many families made how many claims? So, it turned out that 22 families did not make any claim, 53 families made 1 claim, 58 families made 2 claims, 39 families made 3 claims, 20 families made 4 claims, 5 families made 5 claims, 2 families made 6 claims, and 1 family made 7 claims; that means, no family made more than 7 claims over a period. Over a given period, may be say 5 years time or may be over a 10, 10, etcetera.

So here, we want to test whether the, we want to test whether a poisson distribution fits the data appropriately. Now, here we do not even know the parameter lambda of the poisson distribution. So firstly, we will estimate that. So, first we estimate the parameter lambda of the poisson distribution. Now, we have already done estimation. We may use say maximum likelihood estimator or say minimum variance unbiased estimator for lambda. So, for example, maximum likelihood estimator for lambda is x bar. So, x bar is the mean which can be evaluated from here. That is 0 into 22 plus 1 into 53 and so on, plus 7 into 1 divided by 200. The total number of families is 200. So, this value turns out

to be 2.05. So, we may approximately take 2 as the lambda value and we like to check whether this data follows poisson at the rate 2. Now, this is a reasonable approximation because, 2.05 is a value and here, we are talking about the number of claims. So, it is appropriate to take lambda to be an integral approximation of the value and it is extremely close. So, this is fine.

(Refer Slide Time: 36:51)



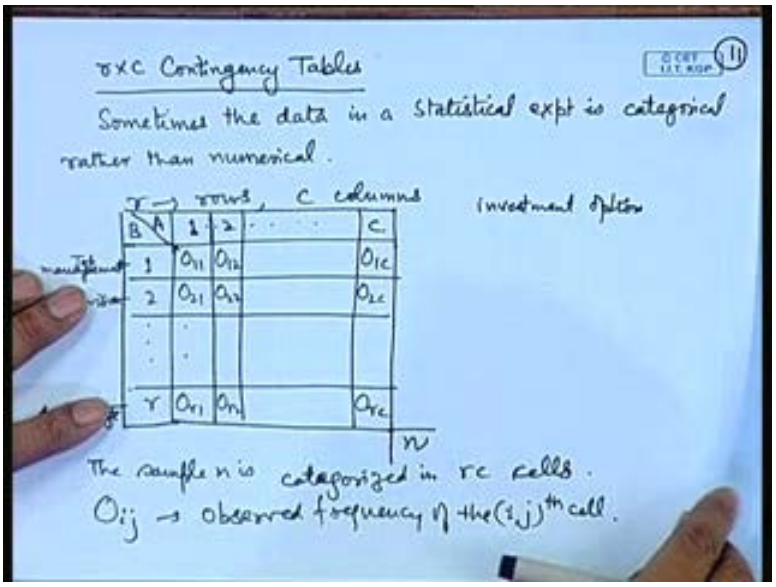
So, we want to find out the probabilities of x is equal to, say small x for 0, 1, etcetera. So, now, the formula in the poisson distribution is e to the power minus lambda, lambda to the power x by x factorial, x is equal to 0, 1, 2 and so on. Now, the point here is that, this is an infinite value distribution. Here, we will calculate the probabilities only for 0, 1 to 7. So, they will not add up to 1. So, what we will do, we can calculate the probabilities 0 1 to upto 6, and 7 we will put into the form of 7 and above. Although, the data is not collected like that, it was observed that for 8, there is no family which made 8 claims, there is no family which made 9 claims, etcetera. So, basically the values 7 corresponds to 7 and above. In that case, the probability of the events, each of the classes will be adding up to 1. So, we calculate the probabilities and we report it like this. So, p i's for 0 1, 2, 3, 4, 5, 6, 7 and above.

So, substituting the value of lambda as lambda head is equal to 2, we can calculate, this is e to the power minus 2, that is equal to 0.135. The probability of x is equal to 1, that

will become lambda e to the power minus lambda. So, that is twice e to the power minus lambda. So, it is 0.27, this value is 0.271 again, because this is again 2 e to the power minus 2, this is 1.8, 0.09, then 0.036, 0.012, 0.005. The total is 1 here. So, now the expected frequency is, the estimate of the expected frequencies can be calculated by multiplying by 200 here. So, we will get it as 27, 54.2, 36, 18, 7.2, 2.4, and 1.0. The total is 200. Now, you note here. The sale frequency of sixth and seventh, 6 and 7 and above, these are much below 5. In fact, even if we merge these 2 sales, the expected frequency will be only 3.4. So, then this will not allow us to use the chi square approximation here. So, what we do, we merge the last 3 sales. So, if we merge these last 3 sales and add up, we will get 10.6 as the expected frequency, which is above 5. So, in place of 8 sales, now we have only 6 sales. So, sigma O i minus e i. So, we will have to then merge the observed frequency also here. The observed frequency merged will give 5 plus 2 plus 1, that is equal to 8. So, now, O i minus e i square over e I, 1 to 6 only. This is W star. That is equal to 2.33.

Now, chi square value has to be on k minus m minus 1. So, here minus 1 because only 1 parameter is there k is 6. So, this becomes 6 minus 1 minus 1 that is chi square 4. Now, if we see the value at say 0.05, that is 9.48, etcetera. So, H naught cannot be rejected; that is poisson distribution seems to be an appropriate fit or appropriate model for the data. At least on the basis of the given data, we have no reason to reject H naught that, the distribution comes from a poisson.

(Refer Slide Time: 42:40)

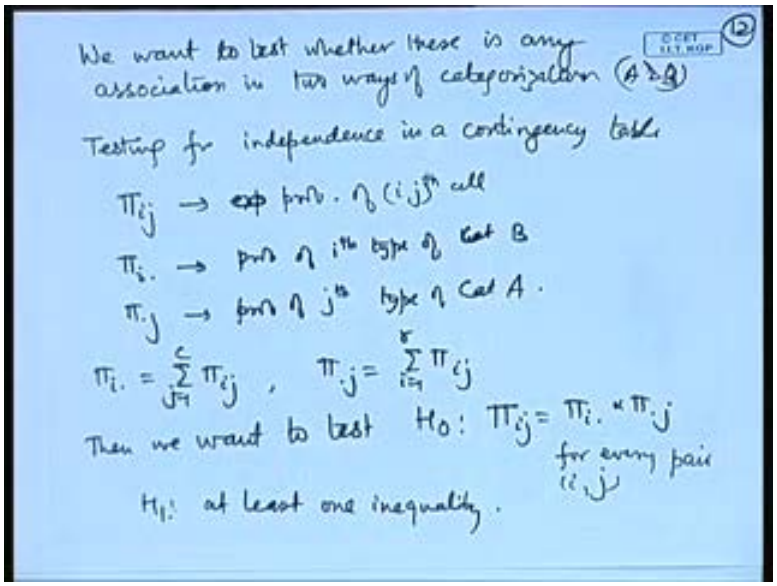


The chi square test for goodness of fit has other applications also. Let us consider the data which is not quantitative, but rather qualitative in nature. Let us consider the situation of contingency tables. r by c contingency tables. Sometimes the data in a statistical experiment is categorical rather than numerical. For example, in a population of individuals, we will like to know how many of the people are smokers and how many of them are non-smokers. So, it is a categorical data; that is, there is two categories of person, one who are smokers, one who are non-smokers. We may also categorize them according to those, who are, who ultimately get lung cancer and who do not get the lung cancer. So, now, the situation is like this in a population, we have characterized according to two different methods of categorization. One is the smoking habit and another is the incidence of a disease.

Now, we want to check whether there is any association between these two characteristics or two attributes; that means, is it true or is it found on the basis of the data, that those who smoke, they get, they are more likely to get a lung cancer? So, this is called testing for independence in a contingency table. So, what is a contingency table? That we have two types of categorizations. One is say, we will represent in the form of rows and another type in the category, in the columns. So, we say r rows and c columns. So, the data may be represented like this. So, we have category A, we have category B. This may be 1, 2 up to c and here, you will have 1, 2 up to r.

For example, in an organization people are classified according to their professional hierarchical levels. For example, one may be the top person of the organization, that is say top management, in 2 we may have some people who are in the supervisory position or executives and here, you may have say daily wage workers. And now, we want to look at the attitude towards a certain option. So, for example, there is a option given for them for investment, investment option; that means, from their salary, they may be allowed to invest in certain, say shares of the company. And there may be, say 4 different types of shares. One is where you get a fixed kind of return, another is where you have a safe return; that means, the equity is distributed into safe instruments, next may be some balanced; that is partially risk and partially balanced, partially safe. And there may be risky option. So, maybe you will find that, out of a sample of n implies of the organization, you found O 1 1 are of, that is those who are in the top management and those who go for option 1, O 1 2, O 1 c, O 2 1, O 2 2, O 2 c, O r 1, O r 2, O r c. So, this is the data. That is, the sample n is categorized in r c sales and O i j is the observed frequency of the i jth sale.

(Refer Slide Time: 48:12)



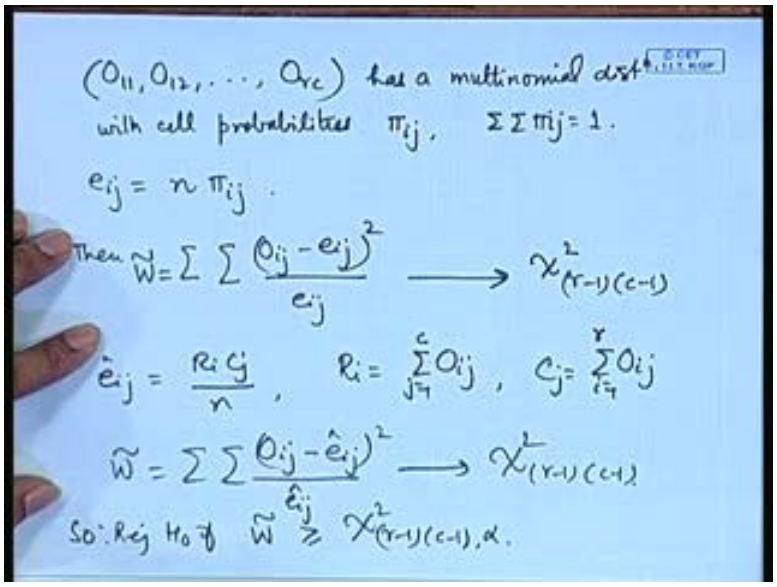
What we want to test here is, we want to test whether there is any association in two ways of categorization, that is A and B. Whether A and B are having any association. As I just gave the example of smoking and cancer, similarly here may be that, those who are

in the higher order in the hierarchy of the company, they may like to invest in high risk equities, etcetera. Whereas, those who are in the lower income group or lower in the hierarchy, they may like to go for safe, this one. Suppose, this could be our hypothesis or we may make a hypothesis that there is no relation.

So, this is called testing for independence in a contingency table. So, you can say here, pi i j is the expected or you can say probability of i jth sale; that is the theoretical probability. pi i dot is the probability of ith type of category, say B. pi dot j is the probability of jth type of category A. So, actually here, pi i dot will be equal to sigma pi i j, j is equal to 1 to c and pi dot j is actually sigma pi i j, i is equal to 1 to r. Then, we want to test H naught pi i j is equal to p i dot into pi dot j. For every pair i j and H 1 is at least 1 inequality.

Now, if we see carefully, this is nothing but a generalization of the test of goodness of fit itself. In the test of goodness of fit, what we wanted to test is that whether the data comes from a particular distribution. The procedure adopted was that, we divided the range of the distribution into k intervals or you can say k sales, and we looked at the observed frequencies. The theoretical distribution of the observed frequencies was multinomial. Now, likewise here, if you see, what we are claiming here is that the probabilities are pi i j's, for the i jth sale the observed frequency is O i j. So, if we look at the distribution of the categories here, that is O 1 1, O 1 2, O 1 c and so on upto O r c. The joint distribution of this will again be multinomial. So, that means, the test will be actually of the previous form itself with little bit modification.

(Refer Slide Time: 52:23)



So, if we write down O 1 1, O 1 2, and so on upto O r c. This has a multinomial distribution with sale probabilities pi i j with sigma of pi i j is equal to 1. So, using the previous argument, if I define say e i j is equal to n times pi i j, then double summation O i j minus e i j square by e i j; that means, call it say W tilde. Then, this is asymptotically chi square distribution on r minus 1 into c minus 1 degrees of freedom.

The values of e i j's are estimated by considering R i C j by n, where R i is actually summation of O i j, j is equal to 1 to c and C j is sigma O i j, i is equal to 1 to r. So, if we make use of this, O i j minus e i j head square by e i j head, then this will be asymptotically chi square on r minus 1 c minus 1 degrees of freedom. So, the test is reject H naught, if W tilde is greater than or equal to chi square r minus 1 c minus 1 alpha.

So, the chi square test for goodness of fit is applicable for testing for independence in a r by c contingency table also. So, in the next class I will give some applications of this test. We will also see that, this is applicable to a slightly different situation. Here, you can see that we are considering a random sample of size n from the population and then we are making this classification. In certain other situations, even this row totals or column totals may be fixed. And then the sampling is done from there; that means, something like a stratified sampling, we will see that even in the stratified sampling the same formula is applicable. So, in the next class I will be covering those portions.