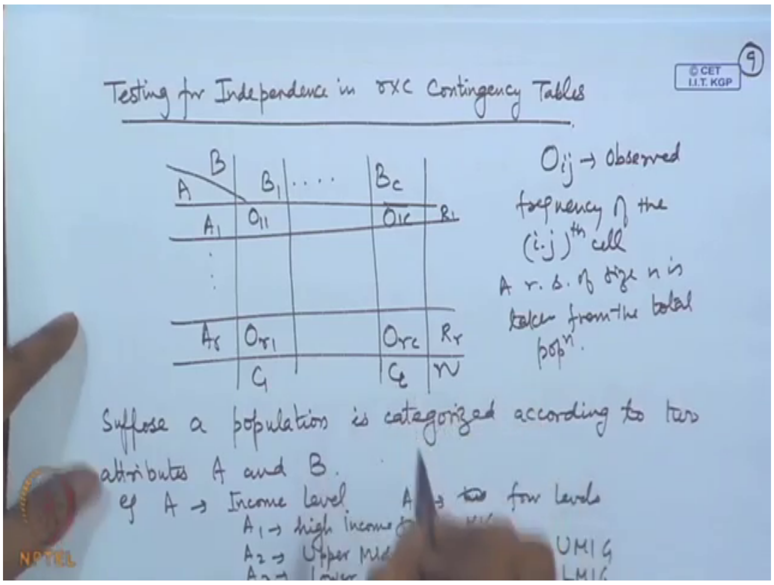
Statistical Inference
Prof. Somesh Kumar.
Department of Mathematics
Indian Institute of Technology, Kharagpur

Lecture – 60
Test for Goodness of Fit – II

If we considered the large sample test actually it is a chi square test. Let me give this thing.

(Refer Slide Time: 00:26)

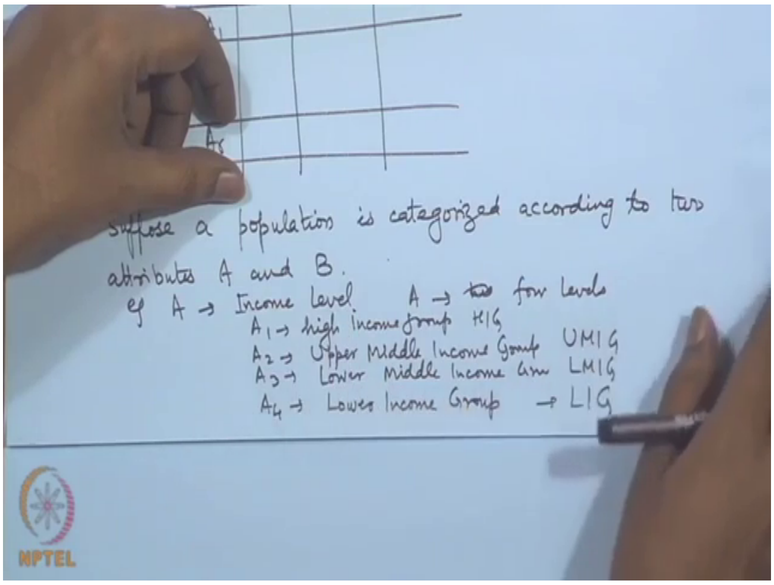


So, testing for independence in r by c contingency tables. So, contingency tables are used when the data is categorical. And so, for example, you may like to test whether the selection of say students in a particular entrance examination dependent upon their financial status, or you can say the financial status of their parents, or is it dependent on the educational status of their parents or is it dependent upon the level of the schools where they are studying the school board and so on.

So, on one side you have certain category let us say A 1 A 2 A r. On the other side you have category B which is having say c classifications. So, this is attribute A this is attribute B. Suppose a population is categorized according to 2 attributes A and B. Say for example, A is the income level. So, we may say A has say 4 levels. Say A 1 is High Income Group say HIG. A 2 is say Upper Middle Income Group that is say UMIG then a

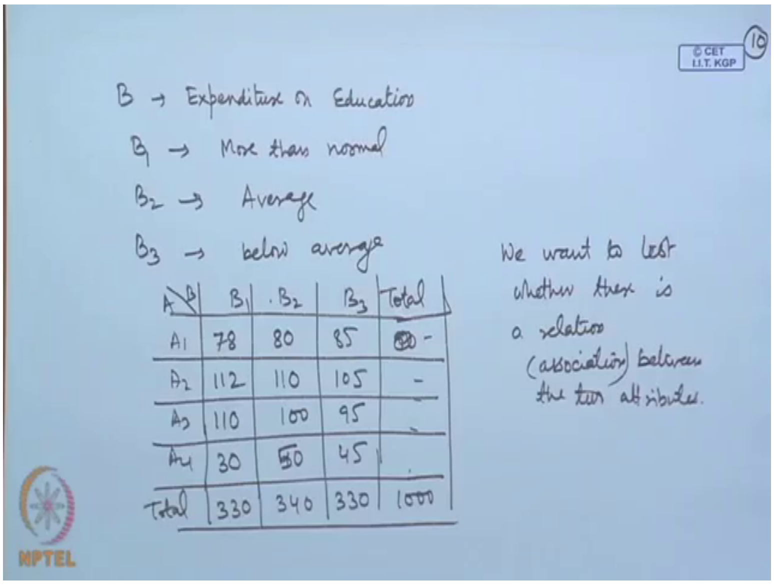
3 say Lower Middle Income Group L M I G and say A 4 is lower income group that is say L I G.

(Refer Slide Time: 03:05)



So, I am distributing or you can say dividing the population into 4 groups according to the attribute which is income level.

(Refer Slide Time: 03:27)

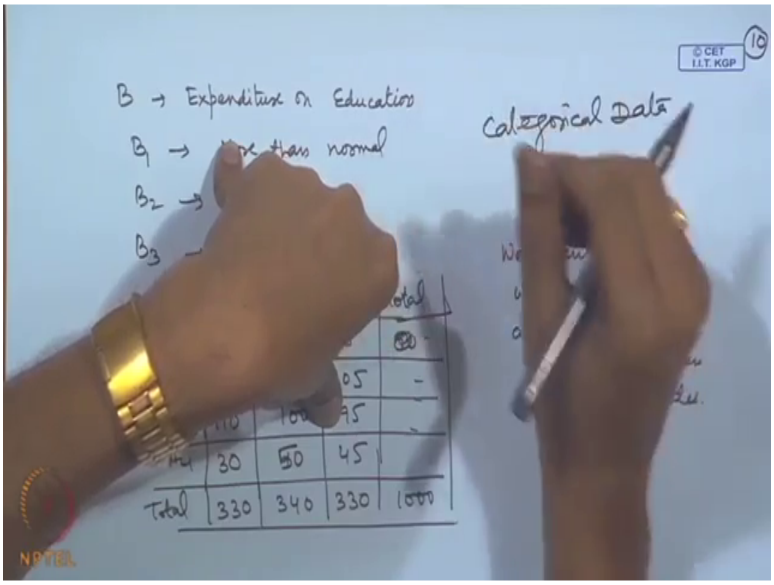


And second one is say expenditure on say education. And then again I put B 1 B 2 B 3. So, this is say more than normal average and below average I consider. So, this is A this is B now A 1 A 2 A 3 A 4. On this side you have B 1 B 2 B 3. I consider say a random sample of say thousand individuals. Out of the thousand individuals how many fall into each category. How many fall into each category. So, there are say 12 classes here. I am just creating an hypothetical data here say 78 1 1 2.

So, this is say let me put 3 classes are there. Let me put some rough number here say 330 340 and say 330. This is just incidental I am fixing the values in some way so, that the total match here. So, this is 190 plus 1 1 0 say 30 80 1 1 0 and say 140 say this is 85, 1 0 5 95. So, this is 140 this is 50 here this is say 90 sorry. So, this totals will add up here.

Now, we want to test whether there is a relation or you can say association between the 2 attributes. Can you say that if the income is high the expenditure on the education will be more? Or if the income is low then the expenditure on the education will be less and so on; that means, is there a dependence on this. So, we want to test this. You can see this is called a categorical data.

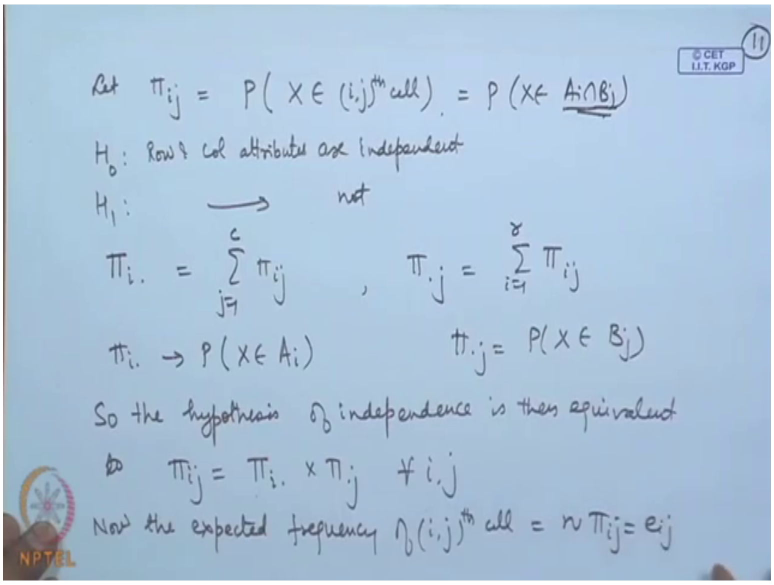
(Refer Slide Time: 06:50)



Categorical data as oppose to the numerical data that you have in the other problems.

Now, for this we can create the situation like this. So, I am making a general setup. You have observed frequencies I will call O 1 1 O 1 2 O r 1 O one c O r c these are the O i j is the observed frequency of the i j th cell. So, this is n here. Now I have considered a random sample of size n. A random sample of size n is taken from the population ok.

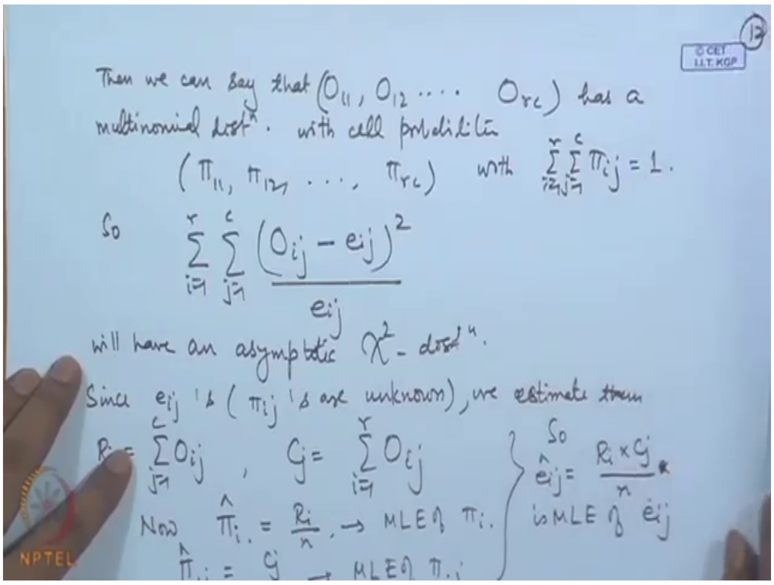
(Refer Slide Time: 07:57)



Let us consider say let pi i j be the probability of the observation belonging to i jth cell. Our null hypothesis is that row and column attributes are independent. And our alternative hypothesis is that they are not independent.

So, we write down the totals here pi i dot that is equal to sigma pi ij j is equal to 1 2 c pi dot j that is equal sigma pi i j over I is equal to 1 to r. So, pi i dot is the probability of observation belonging to the ai. And similarly pi dot j this is the probability of x belonging B j ok. And pi i j is probability of x belonging toA i intersection B j kind of thing. So, the hypothesis of independence is then equivalent to that pi ij is equal to pi i dot into pi dot j for all pi ij. Now the if independence is there expected frequency of i jth cell that will be equal to n times pi i j let me call it e i j ok.

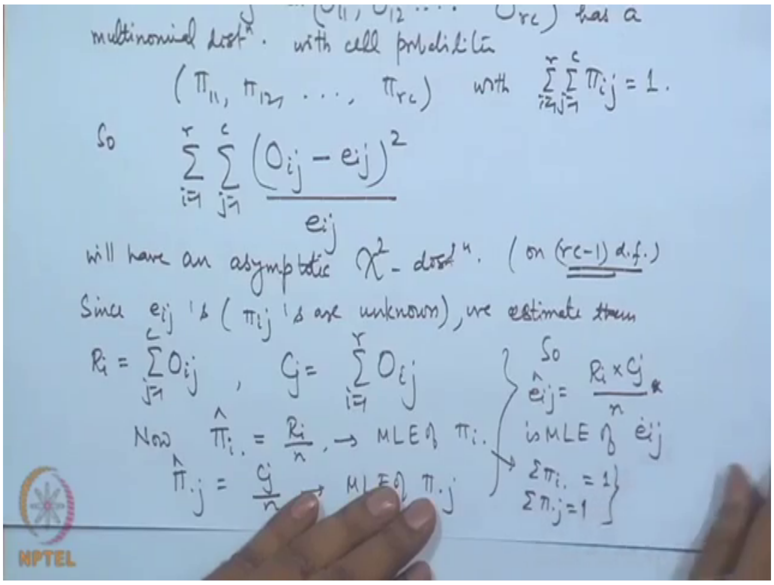
(Refer Slide Time: 10:17)



So, then we can say that O 1 1 O 1 2 and so on all of this taken together this has a multinomial distribution with cell probabilities pi 1 1 pi 1 2 and so on pi r c. And sigma of pi i j i j equal to 1 2 1 r j is equal to 1 to c that is equal to 1. So, by the goodness of fit explanation that I gave earlier, if I considered double summation oij minus eij square divided by eij i is equal to 1 to r j is equal to 1 to c will have an asymptotic chi square distribution. Now how do you evaluate these things? Since the theoretical values are unknown, that is eijs that is pi ijs are unknown we estimate them ok.

How will you estimate that? You can consider say Ri that is equal to sigma Oij, that is the totals on the row side. If I add here I call it R 1 and so on R, r and on this side if I add I call it C 1 C 2 C C. That is a row totals and column totals j is equal to one to C and C j that is equal to sigma O i j, i is equal to 1 to r. Now pi i dot that is the estimate of the ith row probability row attribute probability that can be estimated by R i by n. This is the maximum likelihood estimator of pi i dot. Similarly, if I consider pi dot j hat that is Cj by n ok.

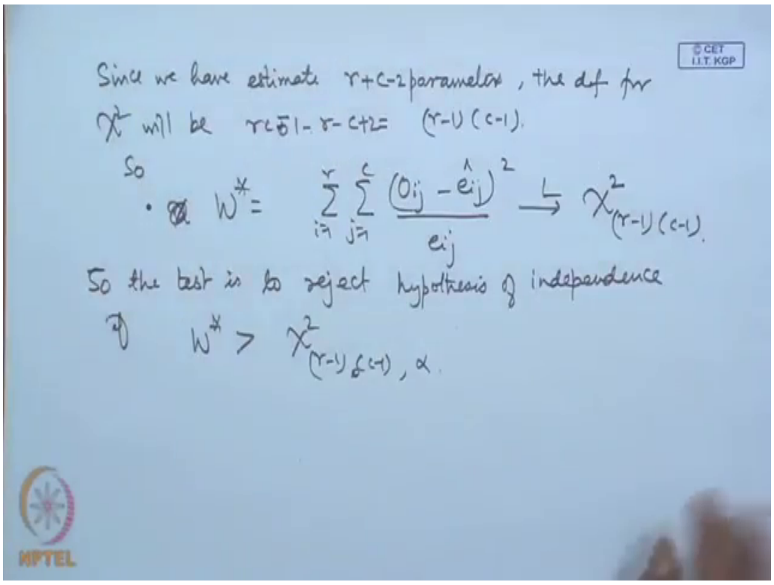
(Refer Slide Time: 13:10)



This is the maximum likelihood estimator of pi dot j.

So, we can consider eij hat as equal to Ri into Cj divided by n into. So, here divided by n and here divided by n and then multiplied by n. So, that n will cancel out I can write just thing. So, this is MLE of eij. So, how many parameters we are estimating here. This total is having rc classes. So, we will have the degrees of freedom r c minus 1 this will have on rc minus 1 degrees of freedom, but now we have estimated r plus c terms.

(Refer Slide Time: 14:17)



Since we have estimated r plus c parameters, the degrees of freedom for chi square will be rc minus 1 minus r minus c. That is equal to r minus 1 into c minus 1 here. I think I have made one error here this will become rc plus 1 minus this one.

So, asymptotic distribution of say let me call it chi W star that is equal to double summation oij minus eij hat square by eij i is equal to 1 to r j is equal to 1 to c. That will be asymptotically chi square on r minus 1 c minus 1 degrees of freedom. This is actually minus 2 here. The reason is that the total is going to be 1. We are getting sigma pi dot is equal to 1 and sigma pi dot j is equal to 1. So, the last 2 parameters are not estimated ok. So, we have estimated r plus c minus 2. So, this is r c minus 1 minus r minus c plus 2 that is r minus 1 into c minus 1 here.

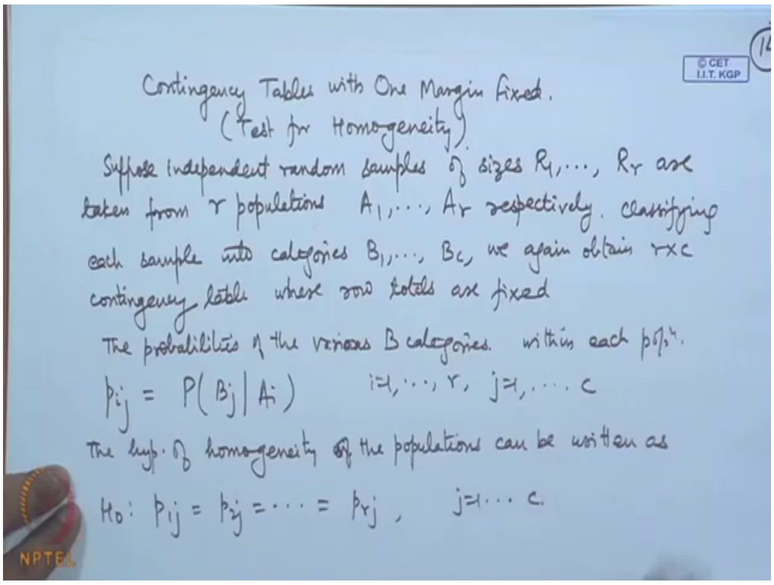
So, asymptotically this will have this distribution. So, the test is to reject hypothesis of independence. If this W star is greater than chi square r minus 1 c minus 1 alpha. Now there is yet either situation here. Because here in this particular case what I have considered the total sample size for the population was fix like n. Therefore, each of this oijs these became random variables. These are also random variables and therefore, they had multinomial distributions.

But there can be a situation for example, the population may be large therefore, we may not fix the total sample. Rather we may fix the sample according to the categories. That is called testing for the homogeneity in a contingency table. It could be like this for example, we may like to we have a distribution of people having different diseases ok.

Now, we categorize the population according to different ethnic groups. So, for example, we are considering say Asians, we may consider say Caucasians we may consider say you say Africans and so on. Now we look at that whether in those populations the distribution of the diseases various diseases is the same or not, but in place of taking the random sample of from the total population, we consider a certain number from say Asians, certain number from Africans, certain number from Caucasians and so on. And then will look at the classification among that. Therefore, now the this ri is are cgs will not be random, but the individual entries will become random.

So, the setup will be slightly modified; however, we will show that the test statistic is still same.

(Refer Slide Time: 18:36)

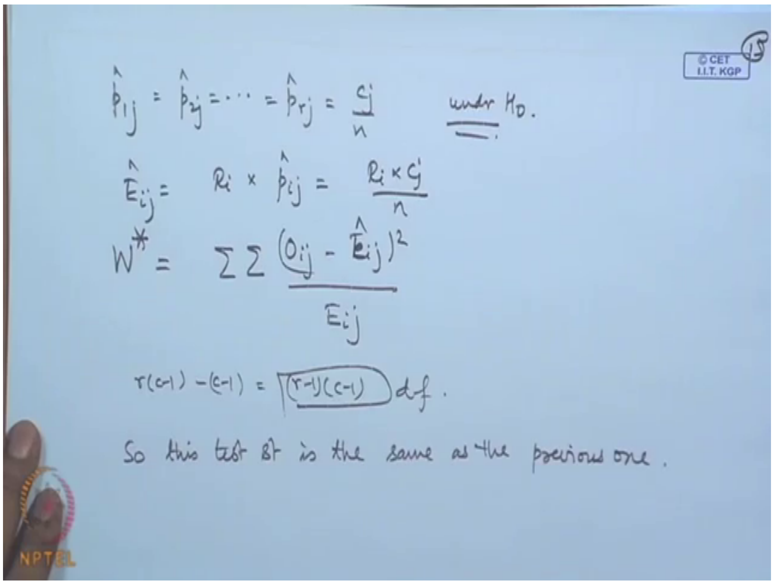


So, contingency tables with one margin fixed this is called test for homogeneity. So, here we are considering suppose independent random samples of sizes R 1 R 2 R r they are taken from r populations say A 1 A 2 A r respectively. In the earlier case I have considered the full population. And then the categories we are called A 1 A 2 and A r and according to the other attribute B 1 B 2 B r. Now here from each of them I am considering the samples. So, this will be called populations and after the sample is taken this is classified into B 1 B 2 that is according to the attribute B. So, classifying each sample into categories say B 1 B 2 B c we again obtain r by c contingency table.

So, the nature is the same; however, the interpretation of the values is different. Where row totals are fixed. Alternatively, you may fix the column totals and then definitely; that means, B 1 B 2 B c may have fixed sample sizes. And then you can classify each of them according to A 1 A 2 A r ok. Now here the probability then will become conditional probability.

So, probability of the various B categories within each population they are now pij that we call probability of Bj given Ai for i is equal to 1 to r, j is equal to 1 to c. And the hypothesis of homogeneity of the populations, this can be written as say p 1 j is equal to p 2 j and so on is equal to p r j for j is equal to 1 to c.

(Refer Slide Time: 22:14)



Now, for estimating this probabilities we can consider p 1 j hat is equal to p 2 j hat is equal to p r j hat that is equal to c j by n. This is under H naught we are considering. So, for i jth cell Eij hat this will become R i into pij hat that is equal to Ri into C j by n. So, it becomes the same value which I obtained in the previous case. And therefore, the test it is statistic is same as W star that is double summation Oij minus Eij hat square divided by Eij. The calculation of the degrees of freedom is slightly different now. You are having c cells and therefore, in each cell you will have each this one you will have c minus 1 degrees of freedom.

So, r times that. So, total degrees of freedom will become rc minus 1. Now in each of them you will be considering the estimated number of parameters that is minus c minus 1. So, that is becoming r minus 1 c minus 1. So, it is the same degrees of freedom. So, this test is statistic is the same as the previous one. So, what we are observing that testing for the independence ina r by c contingency table we are getting asymptotic chi square test. And the test is similar when the interpretation of the or you can say the sampling scheme is slightly modified. In place of taking the sample from the full population I distributed into the stratum and then in each strata I take a fixed sample size.

However, the testing procedure does not get modified. When I discuss the various problems on the testing of hypothesis I will discuss the applications of these 2 procedures which I have described today. In the following lecture I will be discussing another important test that is called a sequential probability ratio tests. So, that I will be covering in the next lecture. I will divot of you lectures and on the problems also for various tests that we have derived in this particular course.