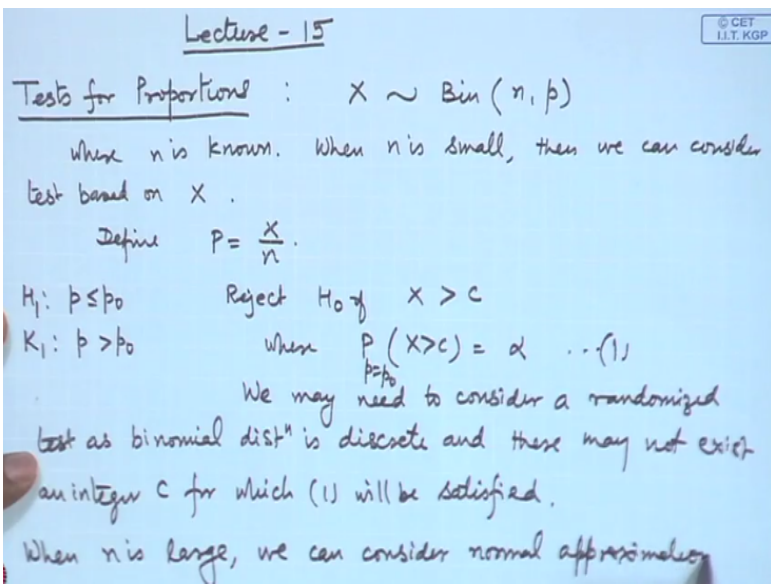
Statistical Methods for Scientists and Engineers
Prof. Somesh Kumar
Department of Mathematics
Indian Institute of Technology – Kharagpur

Lecture - 15
Parametric Methods - VII

We have discussed in detail the tests for the parameters of normal population. I considered one sample problem, in which we considered the testing for the mean and variance of one normal population. We also considered 2 normal populations and we considered various tests for comparing the means and also the variances. However, when we have qualitative data, we may also be interested in testing for the proportions.

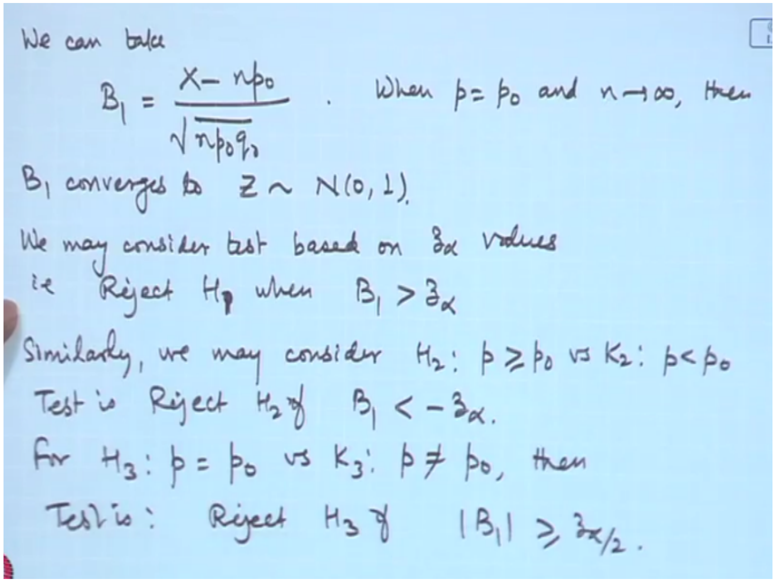
(Refer Slide Time: 00:51)



So here basically the model is that we have x following binomial, say n, p distribution where is known. Now when n is small, then we can consider test based on x. For example, I can call, say let us define say P = x/n. Suppose my hypothesis testing problem is p is </= p0, against p is > p0. Then we can consider the test as reject H0, if x is > some c where probability of x > c when p = p0 = alpha.

Now in this case, what will happen is that it is not necessary that we will get exactly = alpha. So we may need to randomize here. We may need to consider a randomized test as binomial distribution is discrete and there may not exist an integer c, for which one will be satisfied. Now when n is large, we can consider normal approximation.

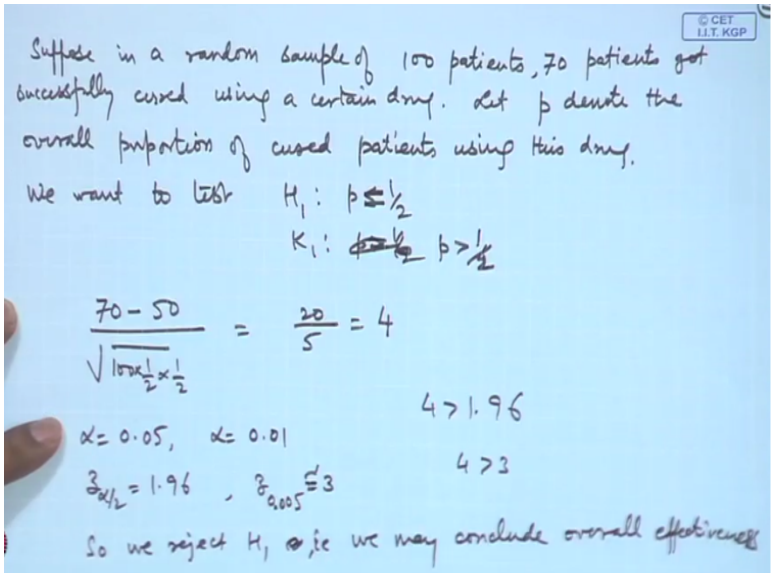
(Refer Slide Time: 03:19)



We can consider x- np0/square root np0 q not. Let us call it say B1. Then p = p0 and n tends to infinity, then B1 converges to z following normal 0, 1 distribution. Therefore, we can for testing about H1 versus K1, for example this hypothesis, we may consider test based on z alpha values, that is reject H0 when B1 is > z alpha. Similarly, we may consider H2 that p is >/= p0 versus k2, p < p0. Then test is reject H2 if B1 < –z alpha.

If I consider the hypothesis p = p0 against p is not = p0, then test is reject H3 if modulus of B1 is >/= z alpha/2.

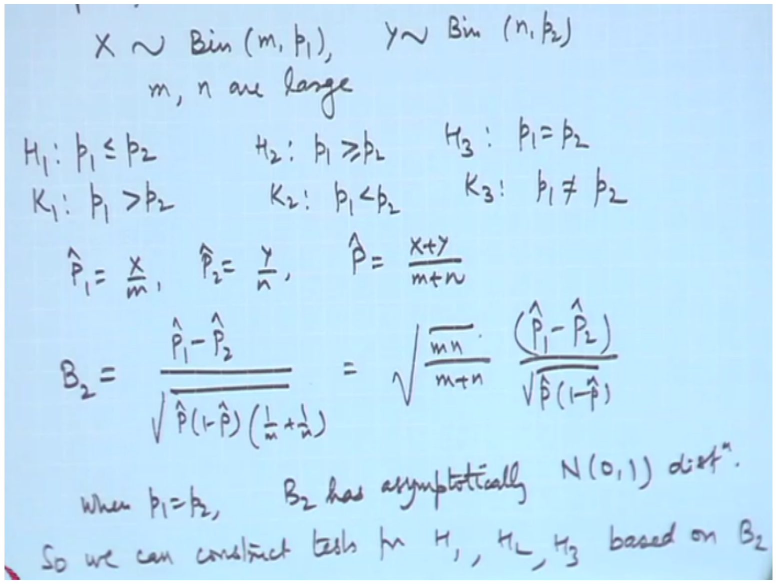
(Refer Slide Time: 05:18)



Let me give a simple example. Suppose in a random sample of 100 patients, 70 patients got successfully cured using a certain drug. Let p denote the overall proportion of cured patients using this drug. We want to test say H1, p <= 1/2 against say K1 p>1/2 or we may say p=1/2 against p >1/2. Suppose we want to test that the overall effectiveness is more than 50%. In that case, the test statistic will become, you will have 70 – 50/root 100 1/2 *1/2, so that becomes 20/5 = 4.

So if I consider say L5 = 0.05 or L5 = 0.01 etc. Then we see that z alpha/2, for example here it is 1.96 and so on. So certainly here 4>1.96. Similarly, at this 1, suppose I say 0.005 then that is still higher value, it is approximately 3, that is > this. We reject H1 that is we may conclude that overall effectiveness of drug is more than 50%.

(Refer Slide Time: 08:22)

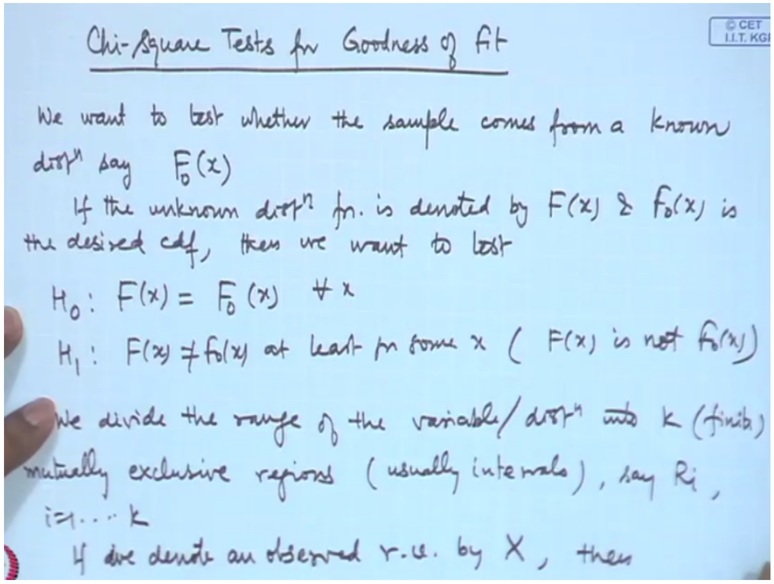


Sometimes we may be interested in comparing 2 proportions. That means we have x following binomial m, p1 and y following binomial n, p2 and n and m are large. We may need to compare p1 and p2, so we can consider hypothesis of the nature this, or say H3. H2 say p1>/=P2 against k2, p1<p2, H3 p1=p2 against k3 is p1 is != p2. So let us refine say p1 hat = say x/m, p2 hat = say y/n, p hat let us define to be x+y/m+n.

And let us define the statistic p1 hat – p2 hat/square root of p hat into 1-p hat 1/m+1/n that is actually = root mn/m+n p1 hat – p2 hat/root p hat * 1-p hat. So when p1 = p2, then B2 has asymptotically normal 0, 1 distribution. So we can construct tests for H1, H2, H3, etc. based on B2. For example, for H1 versus K1 the rejection region will be for z > z alpha. For H2 versus K2, the rejection region will be for z<-z alpha and for H3 versus K3, the rejection region will be for modulus z >/= z alpha/2, if I am considering Leville alpha tests.

Let me also consider another related topic. For example, here we are considering in the binomial 2 categories. So for example, if I am considering one binomial, then it is p and then 1-p as the proportions of the 2 types. Here we are considering p1 and p2. Now in general we can consider more categories, so this gives actually rise to a test called goodness of fit tests.

(Refer Slide Time: 11:43)



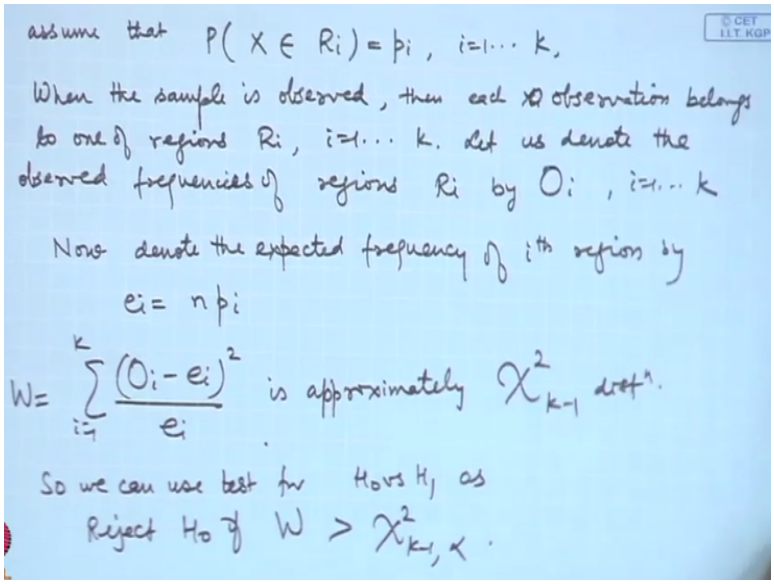
Since asymptotical distributions are Chi square, so the tests are based on that. So we call them Chi square tests for goodness of fit. Let me introduce the problem first. So we want to test whether the sample comes from a known distribution, say F0 x. in the previous problems, in the usual parametric methods, what we are considering is that we are assuming the form of the distribution, like normal distribution, binomial distribution or I have also given the examples of say, exponential distribution or Poisson distribution.

But there can be situations where we would like to test whether we will have a particular distribution, say binomial distribution or uniform distribution or a Poisson distribution, etc. In that case, we will say that the sample comes from a known distribution, say F0 x. So you want to test that, that means if the unknown distribution function is denoted by Fx and F0 x is the desired CDF, then we want to test H0 Fx = F0 x for all x against H1 Fx is not = F0 x at least for some x.

So that means we are saying that alternative hypothesis is that Fx is not F0 x. It could be some other distribution or it may not be distribution. In the Chi square test for goodness of fit, we divide the range of the variable or distribution into k mutually exclusive regions, usually it will be intervals. I mentioned regions, because suppose I am considering binomial distribution, etc., then you have values 0, 1 to n or you are considering Poisson, then it is 0, 1, 2, 3, and so on.

So you will have in finite number of values, but when you can take a practical consideration by considering values by clubbing some of the values together and make it a finite number, so this k is finite. So we divide the range of this, that means we are actually getting some k regions, such as we can give some name here, say Ri i=1 to k and if we denote an observed random variable by x then assume that probability of x belonging to the region Ri is some pi, i=1 to k.

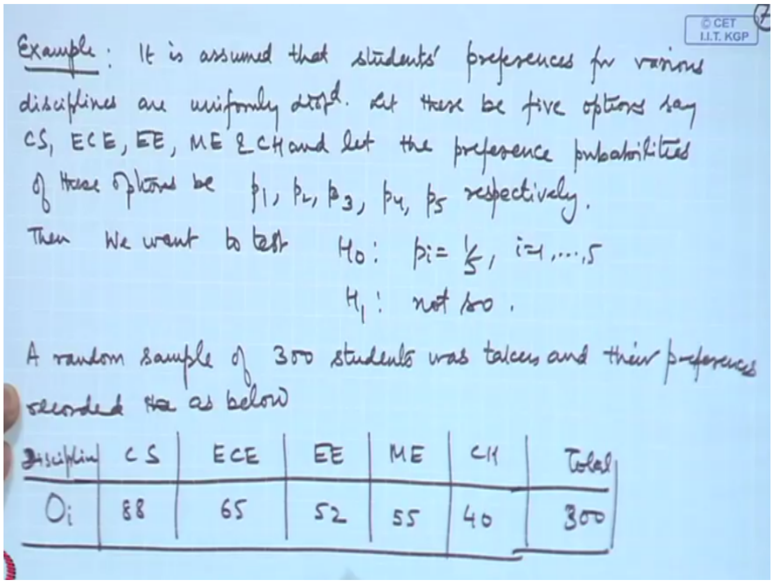
(Refer Slide Time: 15:44)



Now what we consider when the sample is observed, then each xi, each observation belongs to one of regions Ri, i=1 to k. Let us denote the observed frequencies of region Ri/Oi for i=1 to k. So now we consider suppose n observations are there, we denote the expected frequency of i-th region by Ei=n*pi. So what we do, we construct sigma Oi – Ei square/Ei, i=1 to k. This let us call it W, then this has approximately Chi square distribution on k-1 degrees of freedom.

So we can use test for H0 versus H1 as reject H0 if W is > Chi square k-1 alpha. Let us consider an example here.

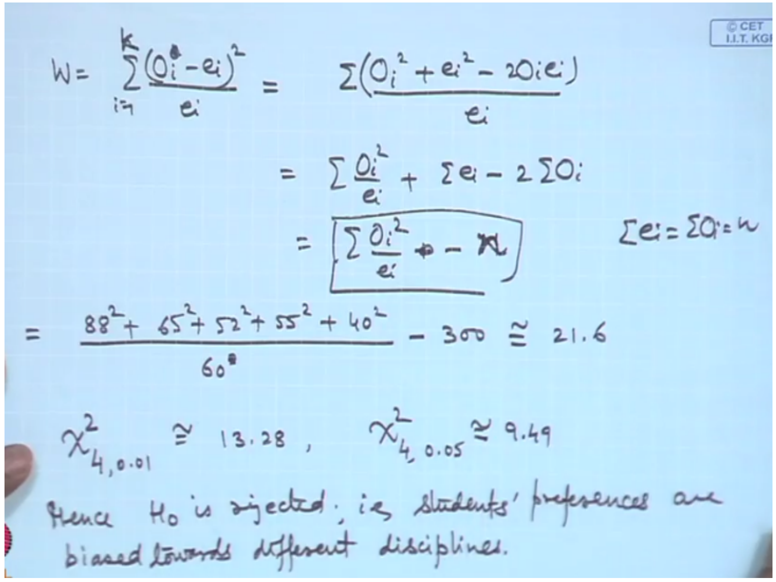
(Refer Slide Time: 18:29)



It is assumed that student's preferences, it is assumed that student's differences for various disciplines are uniformly distributed. So let there be 5 options say CS, EC, EE, ME and CH and let the preference probabilities of these options be say p1, p2, p3, p4, and p5 respectively. Then we want to test, that is pi = 1/5 for i=1-5 against not so. That means we are assuming the discrete uniform distribution for the preferences.

Then a random sample of say 300 students was taken and their preferences recorded as below. So here we have the branches and the observed frequency Oi is given by 88, 65, 52, 55, and 40. So we want to test whether the preferences are uniformly distributed or not. So we consider here Ei's. Ei's are the probabilities of each group. So you notice that, the expected frequency of each group, so if total number is 300, we are assigning probability 1/5 to each group. So the expected frequency will be 60.

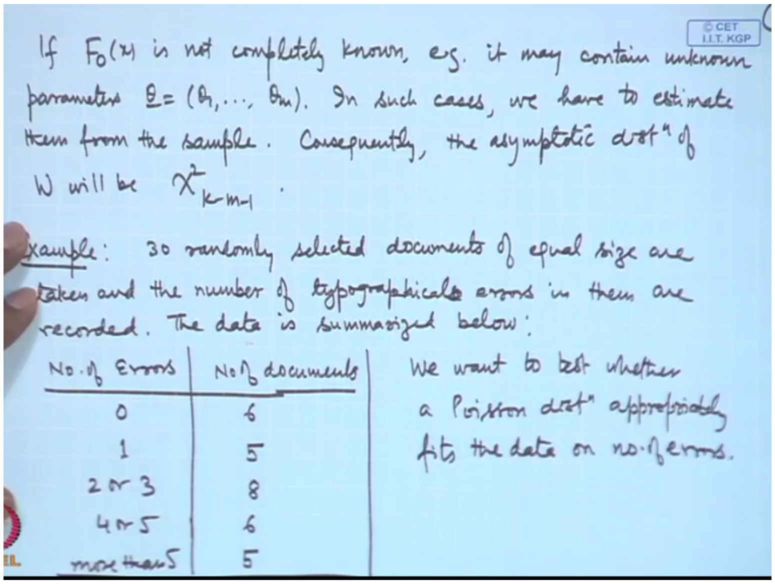
(Refer Slide Time: 22:02)



So we consider here W that = sigma Oi – Ei square/Ei, i=1 to k. this is also having an alternative representation. If I expand this numerator, I get Oi square + Ei square – 2Oi, Ei/Ei that is = sigma Oi square/Ei + sigma Ei-twice sigma Oi = sigma Oi square/Ei + -N, because sigma Ei and sigma Oi both equal to the total sample size. So this is an alternative formula for this, so we calculate here by 60-300. So you can do the calculations, it turns out it is = 21.6.

Now there are here 5 groups, so we need to look at Chi square value on 4 degrees of freedom. For example, we may consider say at 0.01 level, then it is 13.28, suppose we consider Chi square value at say 0.5, then it is equal to 9.49. So you can easily see that H0 is rejected. That student's preferences are biased towards different disciplines.

(Refer Slide Time: 24:27)



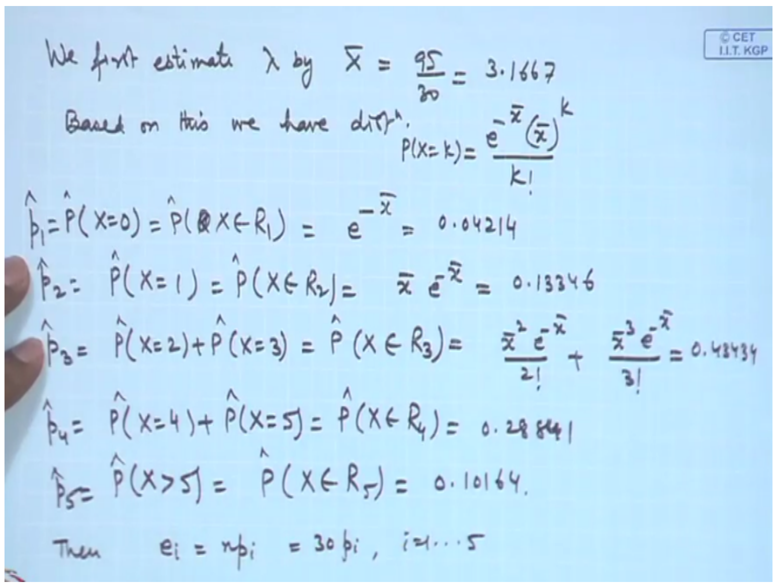
In this particular case, I assume that F0 completely known. If F0 is not completely known, for example it may contain, for example I say it is binomial distribution, then there will be an unknown parameter p, which has to be estimated. Suppose, we say it is a Poisson distribution, then the parameter lambda has to be estimated. Suppose we say it is normal mu sigma square distribution, then mu sigma square have to be estimated first and then they have to be used in the calculation of the expected frequencies.

In that case, the degrees of freedom of the Chi square will be reduced by the number of unknown parameters that have to be estimated from the sample. So it may contain unknown parameters, say theta = theta 1, theta 2, theta n. In such cases, we have to estimate from the sample. Consequently, the asymptotic distribution of W will be Chi square k-n-1. Let us take one example here.

30 randomly selected documents of = size are taken and the number of typographical errors in them are recorded. The data is summarized below. So if I make a frequency table, number of errors, it is recorded like this 0, 1, then 2 or 3 errors, 4 or 5 errors and more than 5 errors. Then, number of documents who had no errors, it was found to be 6, number of documents which had 1 error were 5, number of documents which had 2 or 3 errors was 8.

Number of documents which had 4 or 5 errors were 6 and the number of documents which had more than 5 errors were 5. We want to test whether a Poisson distribution appropriately fits the data. Because here it is a number of counts, errors or counts. So the data on number of errors, now naturally if we assume, so we have to assume a Poisson lambda distribution. Assume x that is the number of errors follows Poisson lambda distribution. Then this lambda has to be estimated first from the given data.

(Refer Slide Time: 29:01)



So we consider this, we will first estimate lambda. So we may consider say maximum likely you would estimate Ru Mu or the method of moments estimator. In the case of Poisson distribution, all of them are the same. It is simply x bar. So here you can see it will be equal to simply 95/30=3.1667. Now based on this, we have distribution written as E to the power – let us call it x bar to the power k/k factorial. That is the probability of x = k.

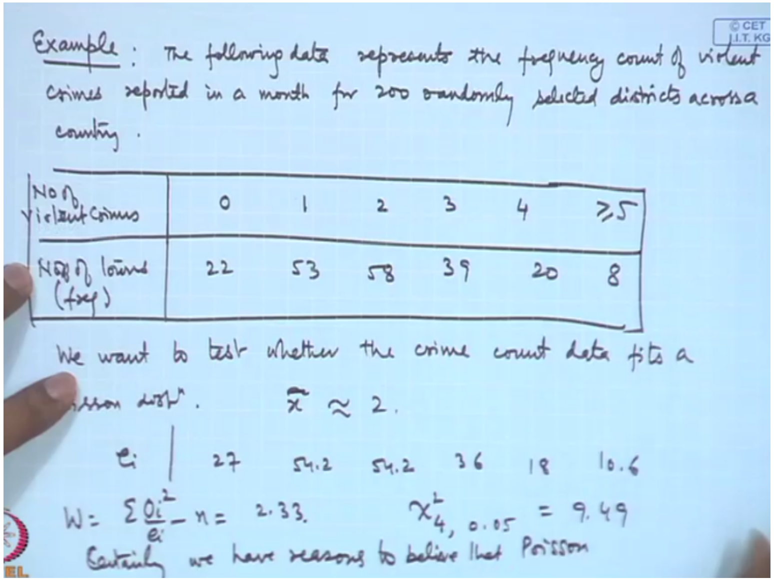
So now for example what is probability of x = 0. See these are the groups here, like I mentioned here in the very first one that this one that we divide into k mutually exclusive regions here. So k mutually exclusive regions here will correspond to, this is region 1, this is region 2, this is region 3, this is region 4, and this is region 5 here. So what is the probability of region 1, that is probability of region 1. What is the probability that x belongs to region 1.

This is my P1, so that is = E to the power – x bar, which of course can be calculated to be 0.04 to 14. Similarly, we can calculate P2 that is probability of x = 1, that is the probability of region 2 = x bar * E to the power –x bar. One can evaluate it; it turns out to be 0.13346. Now P3 will be the probability of x = 2 + probability of x = 3, that is the probability of third region, that is = x bar square E to the power –x bar/2 factorial + x bar cube E to the power –x bar y/3 factorial that is = 0.4343 etc.

Similarly, probability x = P4 that is = x = 4+probability x = 5, that is the probability of 4th region, so that will be turning out to be point 28841. That is probability x = 5, that is the probability of 5th region that is = 0.10164. Now based on this, we can calculate Ei's are nothing but npi that is = 30 * pi, i=1 to 5 and we calculate then. So this I can call P1 hat, P2 hat, P3 hat, P4 hat because these are the estimates of the probabilities of these regions. These are the estimates here.

So we calculate W that will be = sigma Oi square/Ei–n = 21.99. Now Chi square value, you can see, so how many degrees of freedom will be there. We have 5 classes and 1 parameter has been estimated, so it will be 3, so one can easily check the values at some particular level of significance. For example, even at 0.005, it is 12.838, so H0 is rejected that is the error count do not fit a Poisson distribution. Let me give one more example where Poisson distribution will actually fit the given data.

(Refer Slide Time: 34:19)



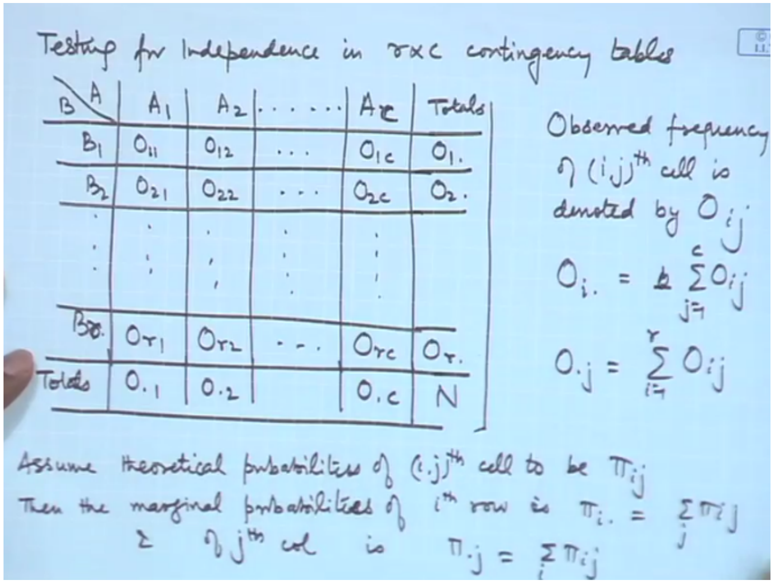
The following data represents the frequency count of violent crimes reported in a month for 200 randomly selected districts across a country. So number of violent crimes and we are clubbing 0, 1, 2, 3, 4, and more than or = 5. So again we would like to test whether it is a Poisson distribution or not. Number of towns, that the frequency, so 22, 53, 58, 39, 20 and 8. So we want to test whether the crime count data fits a Poisson distribution.

So once again, you can check here that x bar is approximately 2, it will be 2. something, so I am just writing 2 here, because that is sufficient for our purpose and we calculate the expected frequencies, expected frequencies will become 27, 54.2, 54.2, 36, 18 and 10.6. So if you calculate W that is sigma Oi square/Ei – n, then that is turning out to be 2.33. So if we look at Chi square value, now since there are 1, 2, 3, 4, 5, 6 groups are there.

The degrees of freedom will be 6-1-1 and let us take 5% level then it is turning out to be 9.49, so certainly we have reasons to believe that Poisson distribution adequately represents this frequency distribution. Now if we see this thing, the fitting of a distribution problem is basically reducing to a sort of multinomial problem because you are dividing the entire categorized data into k categories.

Now if we are dividing into several categories, then it is immaterial whether we divide it into 1 dimension or we can go for higher dimension also.

(Refer Slide Time: 38:26)

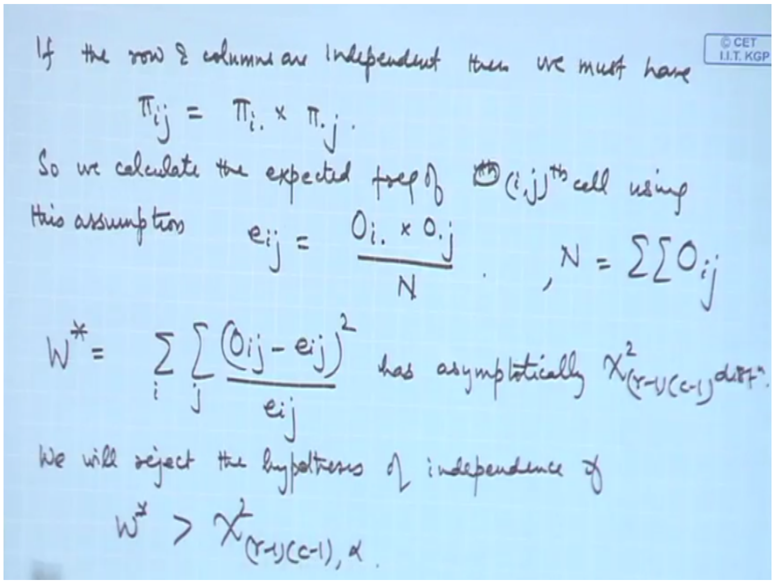


So let us consider in general testing for independence in r/c contingency tables. So if we are considering contingency tables, then we are considering the classification according to 2 categories A and B and for A, we have categories A1, A2, Ar …. Ac and for B we have B1, B2, ...Bc. Now we can actually divide the entire frequency into several cases. Let us put r here. The observed frequencies, I am writing as O11, O12, O1c, O21, O22, O2c and Or1, Or2, Orc.

We consider the row and column sums. So if we sum the first row, we call the sum as O1 dot, O2 dot and so on or dot. Similarly, if we sum the columns, we call that O dot 1, O dot 2, and so on, O dot c. The total sum is n. So we have the following notations observed frequency of ij-th cell is denoted Oij and then we define Oi dot that is = sigma Oij for j=1 to c, so simply the summations and similarly O dot j that = sigma Oij, i=1 to r. These are the row and column totals.

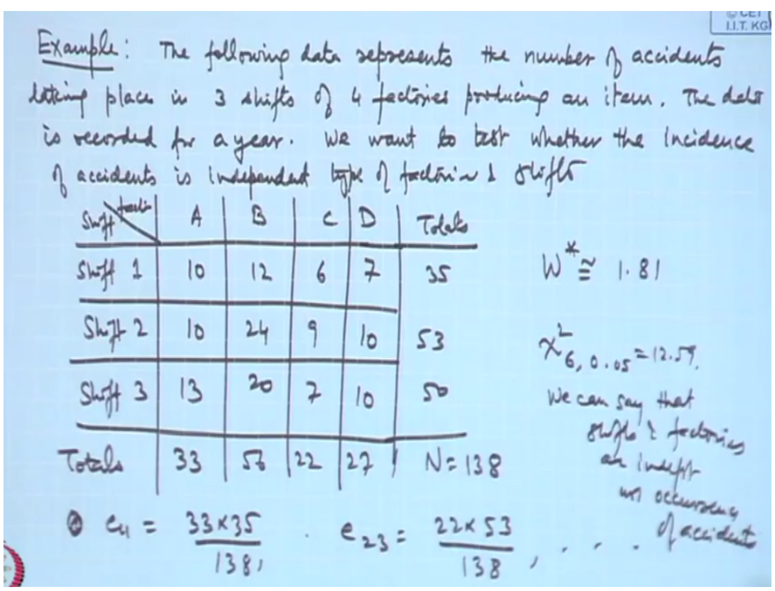
Then if we are assuming that the 2 things are independent, there will be theoretical probability of assume theoretical probabilities of ij-th cell to be pi ij, then the marginal probabilities of i-th row is pi I dot that = sigma pi ij sum over j and of j-th column, it is pi dot j that = sigma pi ij sum over i.

(Refer Slide Time: 42:26)



If the row and columns are independent, then we must have pi ij = pi dot * pi dot j. So we calculate the expected frequency of ij-th cell using this assumption. So that is ij = Oi dot*O dot j/N. Here N is actually the sum of all the frequencies. So if I use this, then we get, let us call it W*=double summation Oij – Eij square/Eij. This has asymptotically Chi square r-1 c-1 distribution. So we will reject the hypothesis of independence, if W* is > Chi square r-1 c-1 alpha.

(Refer Slide Time: 44:40)

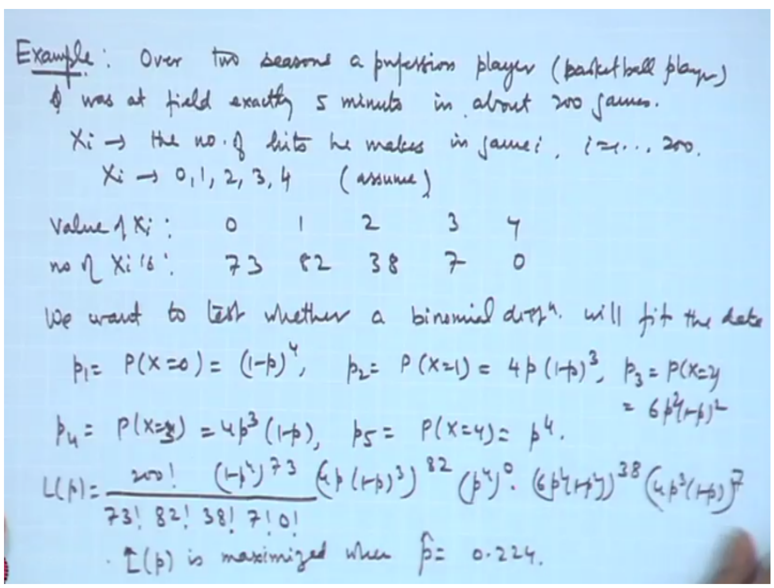


Let me give one application here. The following data represents the number of accidents taking place in 3 shifts of 4 factories producing an item. The data is recorded for a year. So we want to test whether the incidence of accidents is independent, that means whether in a particular factory at particular shift has more accidents are less, so independent of type of factories and shifts. So the data is recorded in this particular fashion.

Suppose we have 4 factories A, B, C, D and the data is recorded over shift 1, shift 2 and shift 3. That is 10, 10, 13, 12, 24, 20, 6, 9, 7, 7, 10, 10. If we consider the totals, this is 33, this is 56, 22, and 27 and on this side, if we consider the row totals, it is 35, 53, 50, the total N = 138. So we calculate for example, what will be E11. E11 will be 33*35/138. Similarly suppose, I consider say E23, so E23 will be 22*53/138 etc.

So we calculate the W* that is turning out to be here 1.81 approximately. Now if I consider Chi square on 2*3 that is 6 degrees of freedom at a particular level, say 0.05, then it is 12.59. So we can say that shifts and factories are independent with respect to occurrence of accidents. You can say that the incidence of accidents is homogenous across the factories. Let us take 1 or 2 more applications of the testing and these problems.

(Refer Slide Time: 49:00)



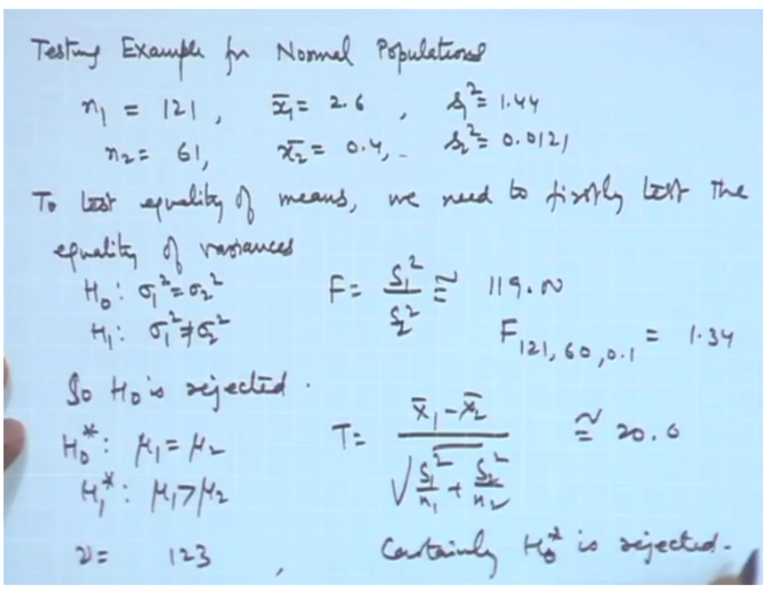
Over 2 seasons, a professional player of some game, we may consider for example a basketball player exactly 5 minutes in about 200 games. So xi is the number of hits he makes in game i, i=1 to 200. Each xi can take value 0, 1, 2, 3, 4. So we have the following data, value of xi is 0, 1, 2, 3, 4 and number of xi is 73, 82, 38, 7, 0. We want to test whether a binomial distribution will fit the data. Now in a binomial distribution, we have a parameter p here.

So let us consider say p hat. Based on this data, we can calculate actually. So P1 that is probability x = 0 that is = 1-p to the power 4, P2 that is probability x = 1 that = 4p*1-p cube, P3 that = probability x = 2 that = 6p square*1-p square, P4 that is probability x = 3 that is = 4p cube*1-p and P5 = probability x = 4 that = p to the power 4.

So we have the likelihood function that is 200 factorial/73 factorial, 82 factorial, 38 factorial, 7 factorial, 0 factorial *1-p to the power 4 to the power 73 4p * 1-p cube to the power 82 p to the power 4 to the power 0 *6 p square *1-p square to the power 38 * 4 p cube * 1-p to the power 7. So this can be simplified L hat p is Lp is maximized when p=0.224. So based on this, we can calculate P1 hat that is 0.363, P2 hat = 0.419, P3 hat = 0.181, P4 hat = 0.035, P5 hat = 0.003 etc.

So if you calculate this, calculate the Chi square value here that = 0.178 approximately. So if you compare with Chi square value on here we have 5 categories and 1 parameter has been estimated, so you will have it on 4 degrees of freedom and one can see this. I will give one application of the general testing problem, which we discussed for the normal populations.

(Refer Slide Time: 54:11)



Testing example for normal populations. The summary data is given by, we have 2 samples for 2 types of elements present in the bones of children and then the following data is collected, n1 is 121, x1 bar = 2.6, s1 square = 1.44, n2 = 16, x2 bar = 0.4, s2 square = 0.0121. We want to test whether the 2 normal populations have similar means or variances. See if you calculate this, firstly we test to test equality of means. We need to firstly test the equality of variances.

So that means, we test say H0 sigma 1 square = sigma 2 square against H1, sigma 1 square not = sigma 2 square. Let us calculate the statistic s1 square by s2 square and it turns out to be 119.00 approximately. So if I consider say F on 120, 60 degrees of freedom, then the values, say at 0.1 that will 1.34 etc. This is certainly larger. So H0 is rejected. So now I consider say mu 1=mu 2 against say mu 1 is > mu 2.

Then we formulate the test statistic x1 bar – x2 bar/s1 square/N1 + s2 square/N2. That is approximately 20.0. If I consider the degrees of freedom of this T here that is turning out to be approximately 123, so certainly H0 * is rejected. So here for testing the equality of the means, which procedure is to be used, because I discussed 4 different procedures, firstly we need to check about variance.

Now for the variance here it turns out that it is rejected here and therefore we have followed this procedure. If it was accepted, then we have to follow another one, which was based on the pooling procedure. So depending upon what actual method will be used, then only you apply the testing methodology. We have discussed some of the important parametric methods. There are many more, but in this particular course, I will restrict attention to this.

In the following lectures, I will move over to multivariate analysis. So we will have elementary discussion of the multivariate normal distribution and then the related distributions and how they are used for certain calculations or computations or inferences when you have multivariate data. So in that following lectures, we will take up that.