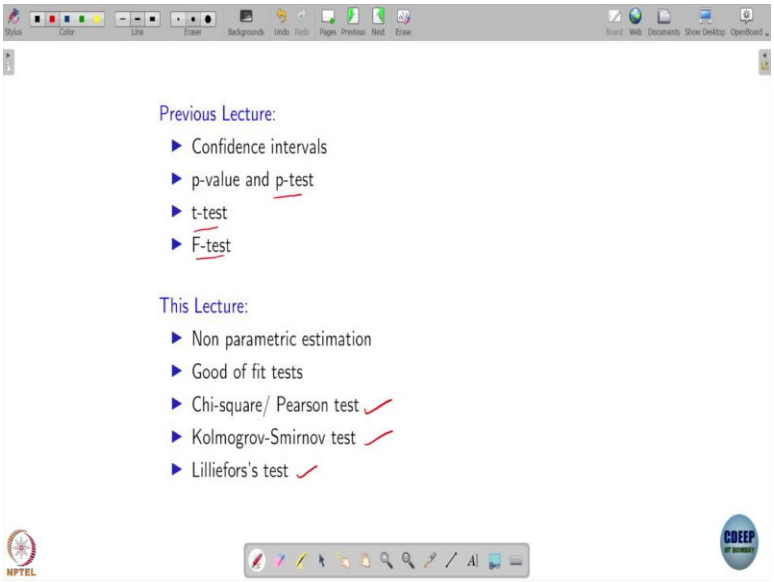
Engineering Statistics
Professor Manjesh Hanawal
Department of Industrial Engineering and Operations Research
Indian Institute of Technology Bombay
Week 11
Lecture 55
Non-parametric test, Goodness of fit, Chi-squared test

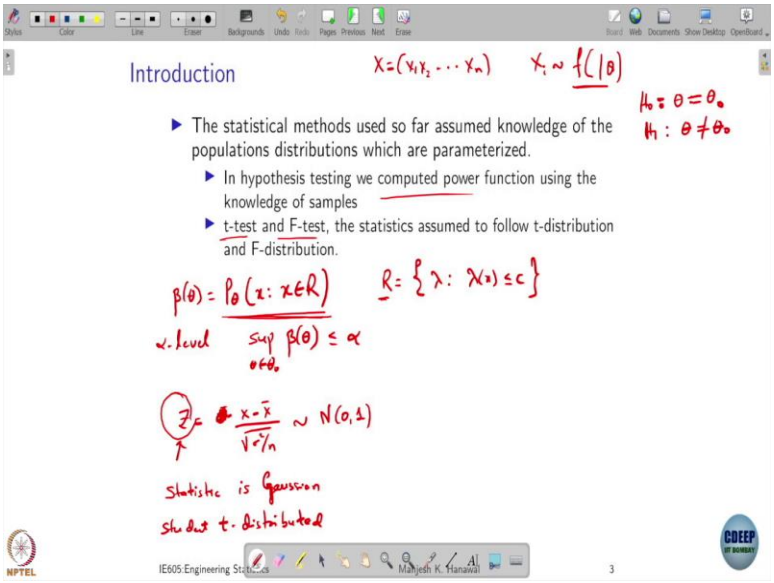
(Refer Slide Time: 00:14)



Hello everyone, welcome to the class engineering statistics again. So, in the previous lectures which talked about confidence intervals, and in particular we focused on how to construct confidence intervals using hypothesis test and then, we talked about various tests based on p values. So, in particular we talked about p values, p tests t test and F test. Now, in this lecture, we are going to continue our discussion of test of a hypothesis.

But, now, we will focus on something called nonparametric estimation that means, the one which does not need to make any assumption about the statistics we are going to use to make our decision either to accept or reject our null hypothesis. And then, we will talk about various goodness of fit test for this nonparametric estimation. In particular, we are going to talk about three goodness of fit test first one is chi square test, then Kolmogrov Smirnov test and then Lilliefor's test.

(Refer Slide Time: 01:44)



Let us get started with what we mean by this nonparametric estimation. So, the statistical methods used so far assume the knowledge of the populations distributions and that to we assume there are parameterized for example, we assume that the sample we have we assumed that this is where Xis are going to come from a certain distribution with a parameter theta and then we set up our tests as whether this theta a certain particular value like is this theta correspond to certain 30 and particular theta not or not.

And we did that using hypothesis testing. And in hypothesis testing, if you recall, we defined something called power function which was like a beta of a theta that is defined as probability that my sample belongs to my rejection region so, where R is the rejection region for example, R could come from your log likelihood ratio test, which defined it like this.

But notice that to compute these probabilities we explicitly need it know this particular distribution and this probability was calculated under the parameter theta and when we try to do alpha level test we try to find something like this belongs to null hypothesis right if you I hope you people all recall this discussions we had.

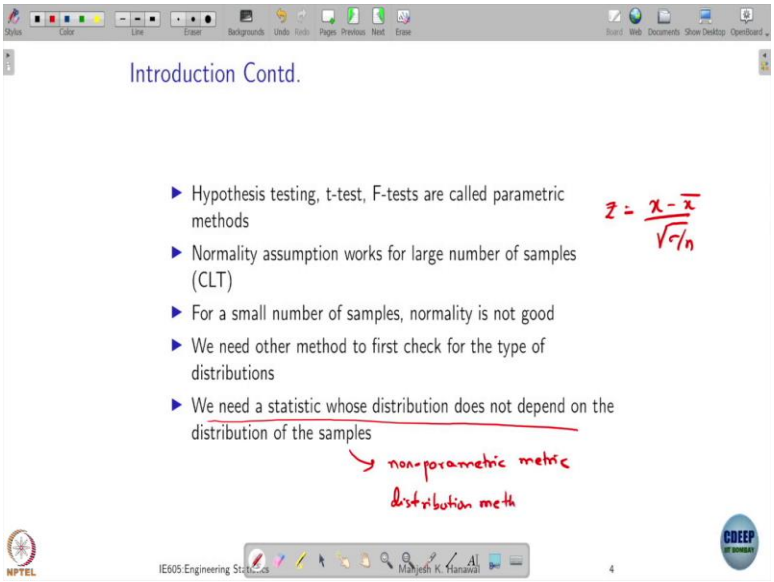
So, in computing all this we kind of explicitly needed to know the underlying distribution and the probability of something falling under this region are we calculated using the knowledge of this particular distribution. And we also looked into various statistics like for example, we use statistics to do t test and F test.

But in doing t tests, for example, if you recall the t test we had something like x bar sorry. I think we had something like x x bar and sigma square and this is we assumed it to be Gaussian distributed, which was the case when it was and the samples are already coming from the Gaussian distribution.

But we assume that this statistic is Gaussian distributed or when the sigma square is not known, we looked into the case when the statistics is student t distributed. So we kind of enforced some distribution on the statistic itself which we use to make our decision. But now, the question is what if we do not want to enforce any distribution on the statistics beforehand that is what if my underlying distributions are not Gaussian, and do I still need to make this make this assumption always to apply this test?

Or put alternatively, if I want to do certain these of this test, I am invariably making this assumption that the underlying samples are coming from Gaussian distribution, but then how to check that indeed the samples are coming from Gaussian distributions for that itself we need a test right. So, thinking all of this we need a method where to know the distribution of the statistics we do not need to know the underlying distribution of the samples itself.

(Refer Slide Time: 06:36)



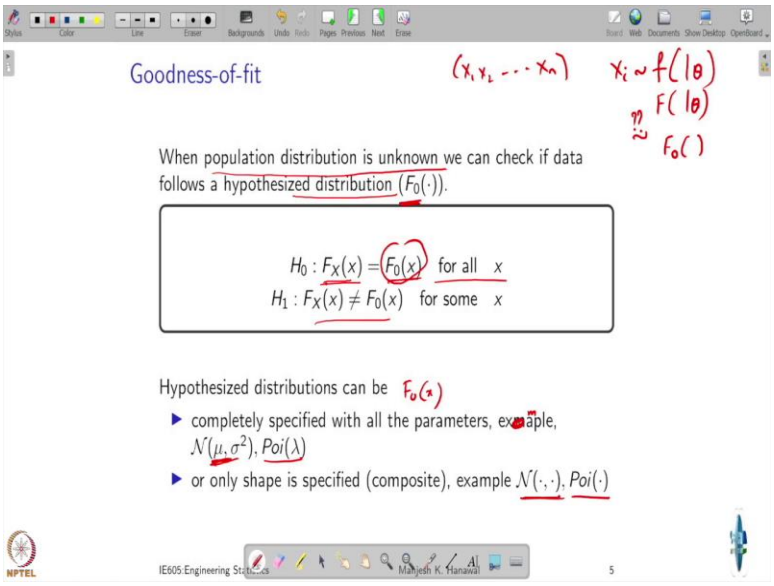
So, keeping this in mind what we have so far discussed is all these hypothesis testing t tests F test these are called parametric method because they explicitly made use of the properties of the underlying distribution or the parameters of the underlying distribution.

And in this at least, the t test and F test, they kind of assumed that normality assumption work holds but this is true only when we have large number of sample and in that case, my statistics which was of the form x bar, which could be think of following Gaussian distribution using our central limit theorem.

But however, this is not the case, when we have small number of samples and that is why before we are going to apply any of the hypothesis testing F test or t tests that we use before we need to validate that the assumptions we are making on the underlying distributions of the sample is correct.

So, to do that itself, we need to have some statistics whose distribution itself does not depend on the distribution of the sample. If that is the case, then we are going to call them as nonparametric method or distribution free methods, distribution sorry, I want to meant here distribution free methods.

(Refer Slide Time: 08:24)



Now, suppose, we want to check whether observed samples are going to follow certain distributions, then we need to have certain tests to check whether they follow they given hypothesis, our distribution, which is taken as null hypothesis, and those tests, we are going to now call it as goodness of tests basically, we are going to say that the samples we are going to observe are they going to follow a given distribution. And we want to check the goodness of that fit.

Now, in that regard, like assume that are like let us say your underlying population distribution unknown. And we want to check if data follows a hypothesized distribution, which I am going to denote as F not. So now, the goal itself is like earlier when I had these samples, I kind of assumed f are going to follow certain PDF, or let us say, some parametric CDF.

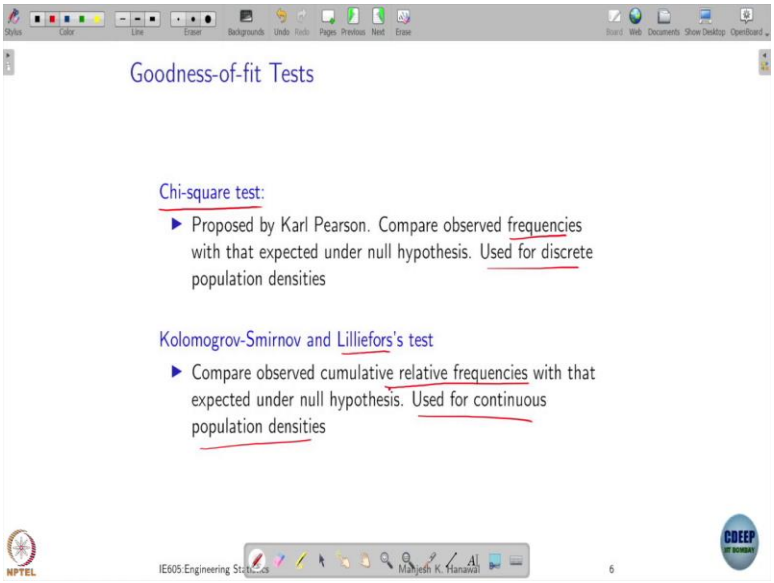
Now I am going to assume that, this itself is not known. I want to check this itself and that I am going to denote it as F of 0 here. Now, then what is my hypothesis now, now, my hypothesis test, my test can be not posed as that my CDF is that of F not, which I am hypothesizing it to be and this holds for all possible values of x.

And the alternative hypothesis says at least it differs at one point that is the my distribution of the data points is not same as the null hypothesis distribution at least one point. Now, this hypothesis distribution now can be this hypothesis distribution can be either completely specified with all parameters.

For example, this F not could be associated with the probability density function, which is Gaussian with parameter mu and sigma squared here the parameters of the distributions are completely specified or it could be told that like our null hypothesis is a Poisson distribution with a parameter lambda or it may happen that this hypothesis distribution is only specified in terms of its shape.

For example, we only know that this is the null hypothesis is a Gaussian distribution. That is it, we do not know, what are the parameters, or we may be just told that it is Poisson distribution without specifying was the parameters is.

(Refer Slide Time: 11:32)

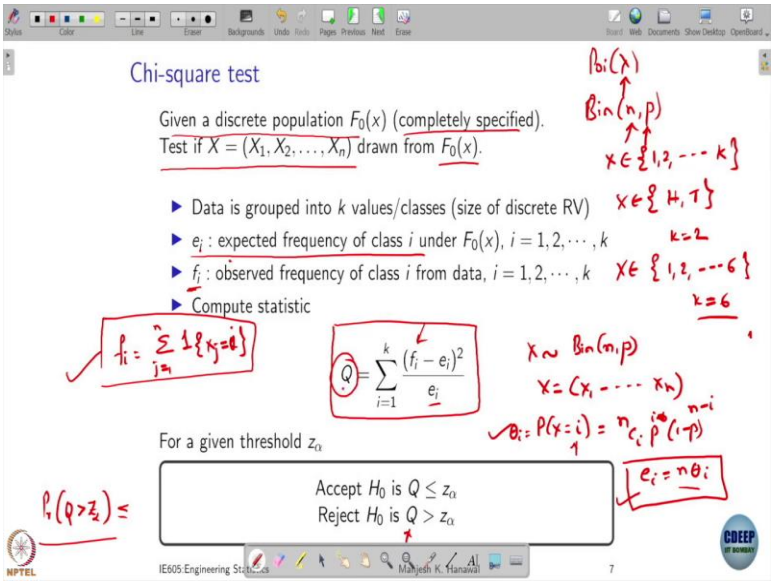


Now, how to go about that how to go about checking whether my samples follow this null hypothesis for that we are going to see majorly two tests one for discrete random variables and another for the continuous random variables in the discrete random variables, we are going to use something called chi square test, which is proposed by famous mathematician Karl Pearson in I think early 800.

So, in this what we are going to do is compare the observed frequencies with that of the expected frequencies under null hypothesis. And as I said, this will be used mostly for the discrete populations. And another test again introduced by famous mathematicians and also statisticians Kolomogrov Smirnov and variant of that by Lilliefor's. So, here we are going to compare observed cumulate you relative frequencies with that expected under the null hypothesis.

So, notice that here we are trying to compare the frequencies of the classes I will make it a bit clear and here we are going to compare the cumulative relative frequencies of the distributions. So, we are going to compare cumulated relative frequencies. And this Kolomogrov Smirnov and Lilliefor's test usually is going to apply it for continuous population density and more specifically, Lilliefor's will be applied to check whether the underlying population is Gaussian distributed.

(Refer Slide Time: 13:37)



Now, let us focus on the chi square test. Now, our we want to test whether my observed data follows a given discrete population, which I do not ask F not, and we are going to assume that, that is completely specified. For example, if it is a Gaussian, sorry, we are going talking about a discrete here, maybe this, F not my distributions can be Poisson with lambda. So here I am specifying the parameter.

Or it could be, let us say a binomial with parameter n and p were both n and p are specified. Now, in this case, if I am given a data, I want to check whether it follows this CDF, how I am going to do that? To do this I am going to group my data into k classes. What are these k classes? For example, let us say that my random variable x is there, which is going to take values 1 2, up to k.

So then there are k classes or my random variables is that it measures take value like head or tail, head or tail. So in this case, my case is going to be 2, these are the two classes. And similarly, if it is, let us say dice, you have, obviously, 6 faces in which my case is going to be 6 like that. Now, what I will be interested in the expected frequency of the class i, suppose I have this Poisson distribution, let us take Poisson distribution.

Let us say my X is Poisson distributed. Now I know that x is going to take the value x is going to take the value 0 1 2 3, like this or maybe Poisson, let us take maybe at this point, it is easier to work with binomial let us say my x is binomial distributed.

And, I have let us say the samples n samples. And I know that probability that x going to take values from i is going to be n choose i p to the power i minus i, and 1 minus p to the power n minus i let us call this as theta. So, this is the probability of observing value i and then my samples the expected frequency of observing class i, is in this case is going to be n theta that is if I have n samples, the expected frequency of observing the class i is going to be n and i in this place, right?

So that is what I mean by ei here. And now, that fi is the observed frequency, what is the observed frequency like now, let me write this observed frequency here. This fi is basically how many times you have observed maybe I will write it as j or maybe it is to write i but maybe fi is equal to basically j equals to 1 to observed j indicator that you were xj equals to 1.

So, it is basically counting how many times you have observed sorry, this is i here. Now, observe you are a random variable has taken value i out of your n samples. So, this is your observed frequency of class i and this is the expected frequency of your class i under your null hypothesis. So, notice that under null hypothesis, because my null hypothesis is completely specified, I know, what are this theta?

Now comes the statistic. Now, what we are going to do is, we are going to check how far this fi and eis are. So, we are going to look into the difference and take the square of that, that is basically the square difference of that and normalize them by their expected frequency, and then submit over all the possible classes. So this is way in a way this matrix is going to capture how different the expected frequencies and the observed frequencies are.

Now, naturally, once we have this, if this difference is too large, it is kind of indication that maybe these what is my observed sample is not following the null hypothesis. On the other hand, if this difference is small, or this some difference is small, then it is a kind of indication that, maybe it is like it is a kind of indication that my observed samples are following my null hypothesis distribution.

So based on this distribution, one can compute or make a decision whether to accept or reject the null hypothesis. So as we did earlier, again, we can set up a threshold and let us say for a given threshold Z alpha. We are going to accept the null hypothesis if this statistic is going to be less

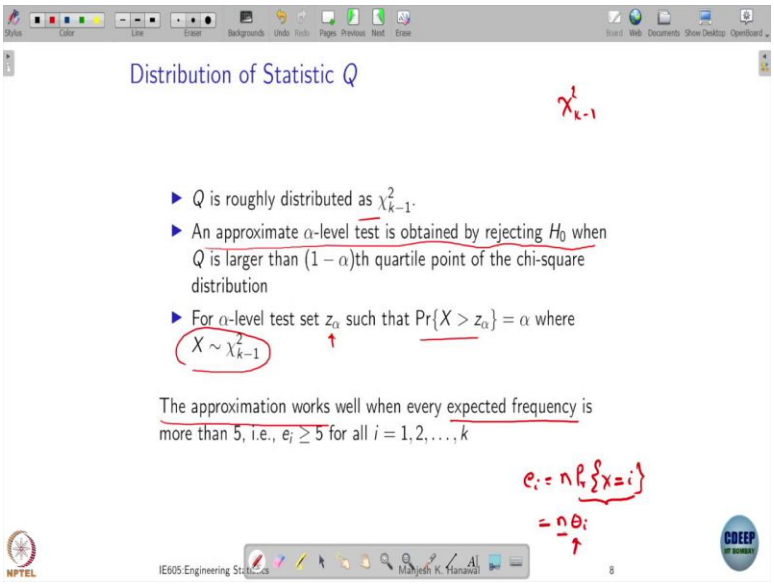
than or equals to Z alpha, and we are going to reject this H0 if it is larger than certain this threshold Z alpha.

Now, the question is can we quantify like how good or bad is our acceptance decision? So, now we may want to compute what is the probability that I reject my sample, can we compute this probability, and that under null hypothesis**.** So, this is we want to quantize and if you are able to say that this is like less than or equal to some number, then that is going to give me the significance of this test with that number.

Now, then the question is how to compute this probabilities? Do we know about this distribution? Earlier, when we talked about hypothesis testing, we kind of assumed this statistic, we assumed its distribution we enforced its distributions to follow Gaussian distribution or to compute its distribution we needed to know the distribution of the underlying samples.

But now, in this case, I am only looking into the empirical values of fi see here q is still a random quantity because this fi are random here. Now, will I be do I need to know the underlying distribution of my samples to compute the distribution or I can say without knowing that.

(Refer Slide Time: 21:46)



It so, happens that in this case, we do not need to know the distribution of the underlying samples. And in fact, one can argue that this Q is roughly distributed as chi squared, this Q has chi square distribution with k minus degrees of freedom. So, we will discuss more about that,

why this is the case, but, notice that without requiring what is the underlying distribution of my samples, I can argue that this Q is going to satisfy chi square distribution with k minus a degree.

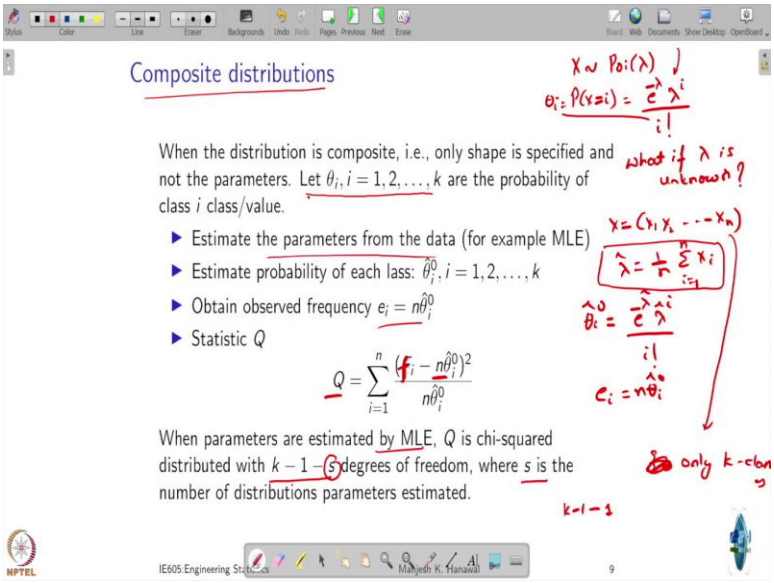
Now, once I have this, I should be able to quantify the significance or the level of my test by setting my threshold appropriately particularly, an approximate alpha level test is obtained by rejecting your null hypothesis when you are Q is larger than 1 minus alpha quartile point or the chi square distribution or like if you are if you want this alpha test, we need to set Z alpha such that you are p of X is greater than z alpha is alpha where you are X is chi square distribution with k minus degrees of freedom and since we know this chi square distribution.

So, well we can compute its tail probabilities and tabulate them and for a given value of alpha, we already know how we should be selecting Z alpha. So, that my test is has a alpha level significance or my test is a alpha level test. So, good now, what we have argued is without knowing the distribution of the underlying samples, we are able to say that my statistic has chi square distribution and we can use readily available table to compute the significance or the level of my test or how should I set up my threshold so, that my test I choose a given significance level.

Now, this test works well when our like this approximation that you are Q is going to follow chi square distribution with k minus of degrees of freedom, well, when the expected frequencies are more than 5 that is when you are ei is greater than 1 for all the classes. This did not be the case all the time, but this is a kind of thumb rule, which will can be used when you have when you want your approximation to be good.

Like for example, we said that ei is going to be n into probability that if your random variable takes value i equals to n. So, whatever this is like theta i this is known under your null hypothesis if you are n is such that it makes your ei larger than 5, then this is a very good approximation.

(Refer Slide Time: 25:21)



Now, composite distribution as I said every time the null hypothesis may not be completely specified, only the shape may be given in that case we call the distribution is composite. And now, how to go about in this case now, in this case, let theta i are the probabilities of the class. Now, because the underlying parameters of the distribution is known, we do not know this theta i exactly.

So, what I mean by this suppose, for example, if X is Poisson distributed we know that probability that x equals to i this case theta i is equals to e to the power minus lambda, I hope I am going to make it correct and this is i lambda but, so, to compute this theta i, I need the knowledge of lambda, but what if I do not know this.

So, in this case, you may estimate this lambda itself from the data that is if you have your samples you can estimate your lambda to be the empirical mean of this data and we know that this is a good estimator it is and unbiased asymptotically it is consistent and, all and then once we have this, we can plug in this lambda i's in this expression here.

And then you may get lambda hat, lambda hat i, and i! and that is what I am going to call it as theta i here based on the estimate. Now, once I have this now, I can also get an estimated frequencies which is now theta i hat. Now, all I need to do is for my given ei values now, I need to compute my statistic, which is same maybe I had to write this as fi, I think earlier I wrote this small fi and I have this value only thing is I have replaced this e by n theta hat 0 subscript i here.

Now, what about the distribution of k, Q does it follow the same distribution like we had earlier like which is a chi square distribution with k minus 1 degrees of freedom. It so, happens that when the parameters that you are going to estimate are based on maximum likelihood estimator, that indeed one can argue that Q still remains chi squared has maintain chi square distribution, the only thing is now, the degrees of freedom is now k minus 1 minus s.

Now, what is this s here earlier it was k minus 1 now, you are saying it is k minus 1 minus s. Now, here s is the number of distributions parameter that we estimated for example, here right in the Poisson we estimated one parameter, in this case, if you have to apply this method, so, then let us say we have observed only some k like even though there are infinitely many classes in this Poisson distribution because k, the X can take value from 0 to infinity.

Let us say you have, in the samples that we have observed in this I only see certain k number of possible classes they belongs to let us say only k classes. So in that case, my k minus 1 and minus one more term 1 is going to come because I have estimated one more parameter so this is going to be k minus 2 in this case. So let us stop here. And then let us continue how is k is indeed chi square distribution with a rough proof sketch.