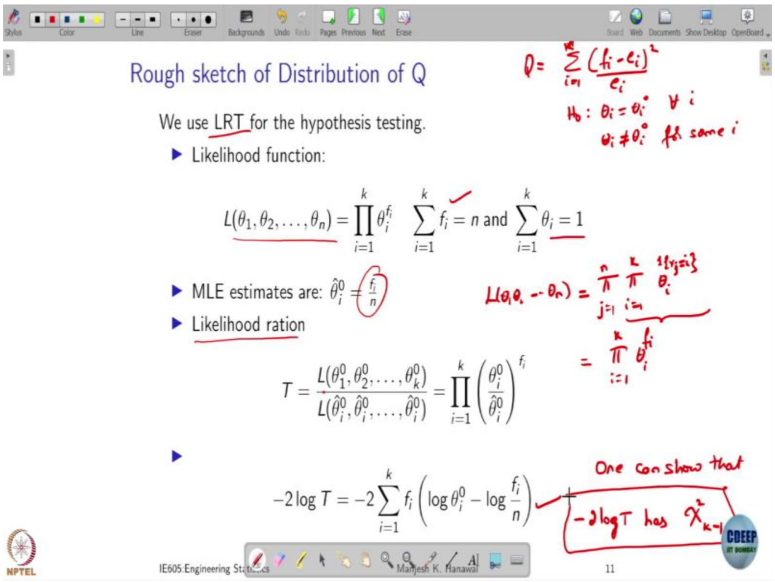
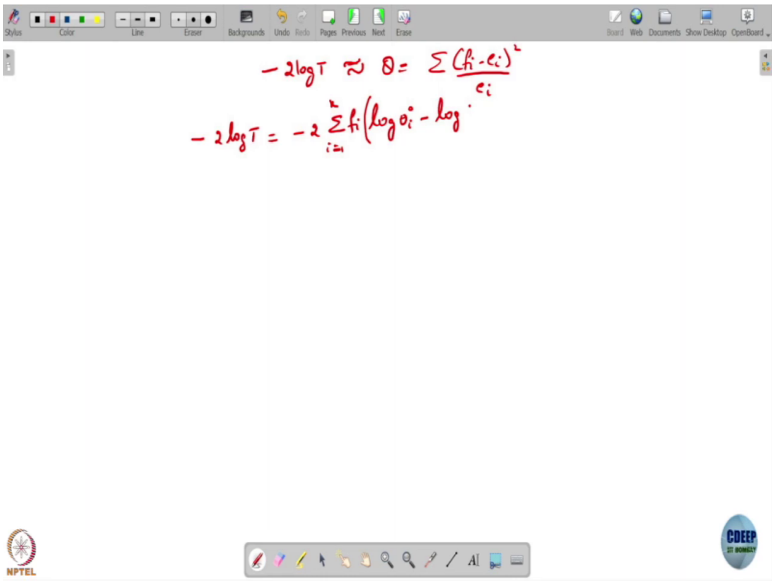
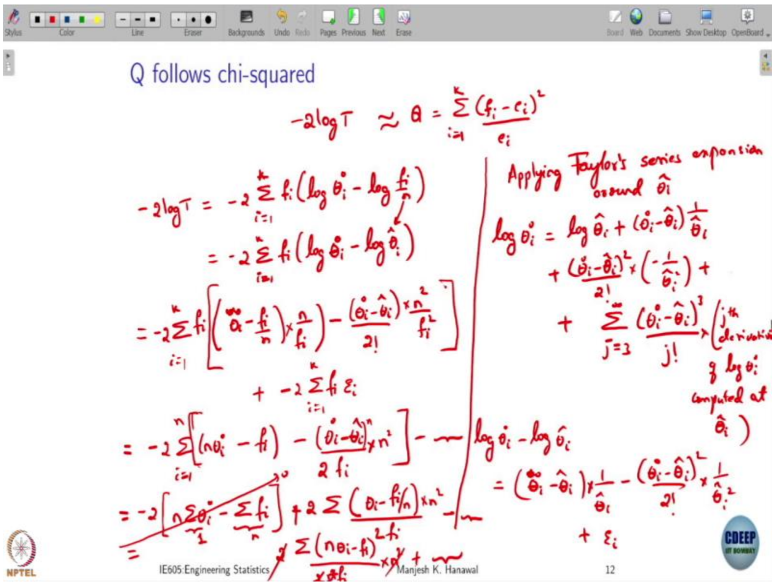
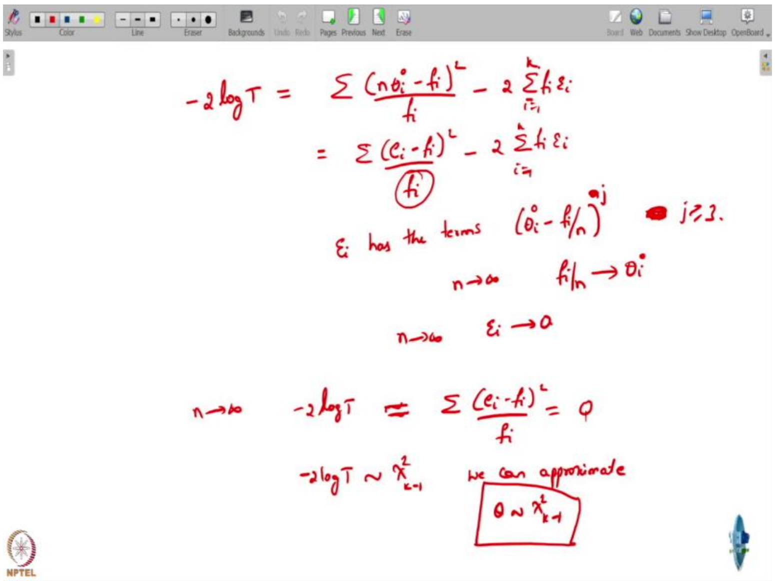
Engineering Statistics
Professor Manjesh Hanawal
Department of Industrial Engineering and Operational Research
Indian Institute of Technology, Bombay
Lecture 56
Distribution of Chi-squared Test Statistics

(Refer Slide Time: 0:23)









Let us get started again. So, now, we wanted to check, what is the distribution of the statistic, we have here like the our statistic is the squared difference between the classes of all the classes between their empirical values, and the expected, empirical frequency, and their expected frequency normalised by their empirical frequency.

So, to this, to do this, sorry this is k here, you are going to use our likelihood, we can use a likelihood test, let us try to check whether my hypothesis is null hypothesis is like a theta i is whether it is equals to theta i 0 for all i, and alternate hypothesis, this is not equal to for some i. To check this hypothesis, I can go and use my log likelihood ratio test.

So, in the log likelihood ratio test this being a discrete valued function, so how am I going to get this, let us quickly write this. So, my L of theta 1, theta 2 up to theta n can be written as this is basically j number of samples, and then i equals to k, that is like a theta 1 indicator that xi sorry, xj equals to 1, so, this is for each sample, I am going to check, and this is going to be i, and this is going to be xj equals two i.

Now, we know that, if I am going to expand this, I will see that theta i is going to be exponented to 1, fi number of times. And that is why I am going to get this quantity to be simply product of i 1 to k theta of fi. But however, I know that there is an this likelihood function, I want under the condition that my summation of fi's has to be and this is naturally attached to hold.

But also, I want this theta i should be such that they add up to 1, because these are the probabilities of the various classes which need to add up to 1. If I use my likelihood like made estimator, one can see that when my optimise this over my thetas under this constraints, I will end up getting theta i has to be fi by n, this is basically the empirical frequency of my class i.

Now, by taking the likelihood ratio test, if you recall, that numerator is going to be simply L computed at my null hypothesis parameters, which are theta 0s, and my denominator is going to be the optimal value of L over my possible parameter space, which is theta hat, L of theta hat. So, by simplifying that, I get this. And now further, if I am going to take minus 2 log of this, I am going to get this expression.

Now, our claim is, or one can show that this Q, sorry, this 2 log T has Chi-square distribution with k minus 1 degrees. So, this is one can show. Now, what we are going to observe is, this minus 2log T we have, this is almost same as Q, which is f i minus e i n squared divided by e i. Now, or, we are going to say this is approximately equals to.

Now, the question is why this is true? Let us rewrite the expression for log T here, my 2 log T, we just saw that this is going to minus 2, i 1 to k, f of i log of theta i, sorry theta i 0 minus log of what is that we have here, I have, I am going to use this slide instead of that. So, we want to show that minus 2log T is almost same as Q, which is summation fi minus ei square, by ei, equals to 1 to k.

Now, let us first write expression for minus 2log T we have, this we have argued that is nothing but minus 2, i equals to one to n, sorry, k, fi log of theta i 0 minus log of fi by n. Let me if we write this minus 2, i equals to 1 to k, fi log theta i n minus log, this I am going to write it as so this is like theta 0, going to write it as theta hat.

Now, there I have just replaced this by this for simplicity. Now, let us try to understand how this expression looks like by using our Taylor expansions. Applying Taylor's series expansion around the point, theta i hat. So, then I am going to get this theta i 0 equals to theta i hat, plus theta i 0 minus theta i hat. And then I need to take the first derivative of this that is, 1 upon theta i hat plus theta i 0 minus theta i hat squared, then two factorial and then the second derivative of this, which is minus 1 upon theta i hat square.

And there will be some additional terms here, which I am going to simply write it as j equals to 3 to infinity, which are like theta i 0 minus theta i hat 3 j factorial times the n time derivative of I am going to just write n time derivative of, and this when I write this is like nth, sorry, jth derivative of your log of theta i.

Look, let us see, what is the nth time derivative of this, jth derivative of this, so I am going to write some, let us say, this is like, I am simply going to write jth derivative of your log theta i 0, computed at theta i hat. So, now, if I simply, so I am going to write this entire quantity now as, if I am going to write it as theta i 0 minus log of theta i hat, this is going to be simply theta i hat minus theta i hat divided by 1 upon theta i hat, plus now there is a minus here, theta is 0, theta i hat square divided by 2 factorial, 1 upon theta hat i square, and this rest of the terms, I am going to simply write it as epsilon.

Now, one thing to notice here is let us get the terms back here. Let me start from here, minus 2. Now, also when I plug in back, I will try to replace theta i by its corresponding quantity, i 1 to k, I have epsilon. And now this is like, theta I hat, but theta i, sorry, this is supposed to be theta 0 here, theta i 0 and theta i hat, I know that this is fi by n divided by this is also then this is going to get it like this, and then minus theta i 0, I am going to keep this as it is, but this one, I am going to write it as n square by fi square, maybe I should write this entire thing like this, divided by 2 factorial. And of course, there is a minus 2 equals to 1 to k, fi and epsilon write down here.

So, now further, let us do this simple computation. And now if you do this, what I am going to get is fi times theta i 0. And this gets this quantity. And I am going to get minus n and n, and I am going to get this fi here, and if you further simplify this quantity here. This is going to be to theta i 0 minus theta i hat. And this one f, this is going to be f and it still has this n square here.

Now, let us quickly compute this, now I see that, I have these two square, I have this, I think I made one competition mistake here, this should be n here instead of f here. So, this is fine.

And this is fine here. So, the only part was this was a mistake here. So, this was like f, f get cancelled, this was like, n theta i not.

Now, if you notice this part, now I am going to expand start expanding this summation by taking inside. So, the first term is 2, then into summation theta i 0 minus summation of fi, and this is the first part. Now, the second part is, so there is something extra, which I am still retaining as it is, minus there is also 2 here. Now, if I see this, one, F has been knocked off one f remains, and I am now going to write expand this, this is going to be a minus, minus is going to be 2, and summation theta i. So, now I am going to write this as fi by n into n square by fi minus still something.

Now, quickly see this. So, what we have, now what we have is basically 2 times log of T is equals to. So, let me write this entire thing, maybe before I go here, let me just let me see some make some space here and write. Now, let us see that this quantity is going to be n, and this is going to be 1, so this quantity gets knocked off.

So, now, if I simplify this, I am going to get it as 2 times summation n theta i minus fi, and n fi into n square plus a term here. So, what we now had is basically let me again just to recap this we have, and also there was a 2 here, so, I am going to do this 2 gets knocked off, and I have in the denominator n square here, because that is what came, and this n square gets knocked off with this and what I will end up is summation n theta i 0 minus fi square divided by times fi n by, so, this one like a whole square here, and I had this square here.

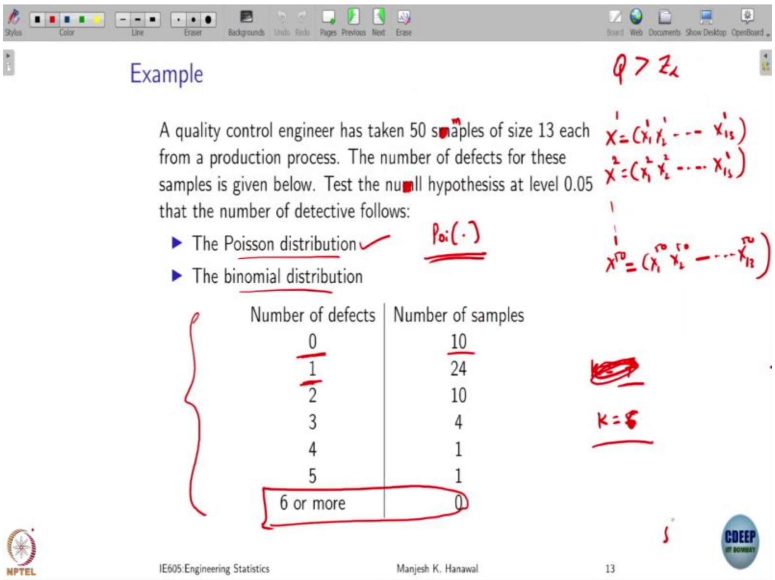
And when I computed this n square is knocked off with this, but in that denominator, I am getting fi, whereas I am anticipating ei in this let me check what is the mistake I made, so as of now, I do not see the mistake here. So, what I have is minus then theta fi plus and I have this minus 2 this fi, epsilon i equals to 1 to k here.

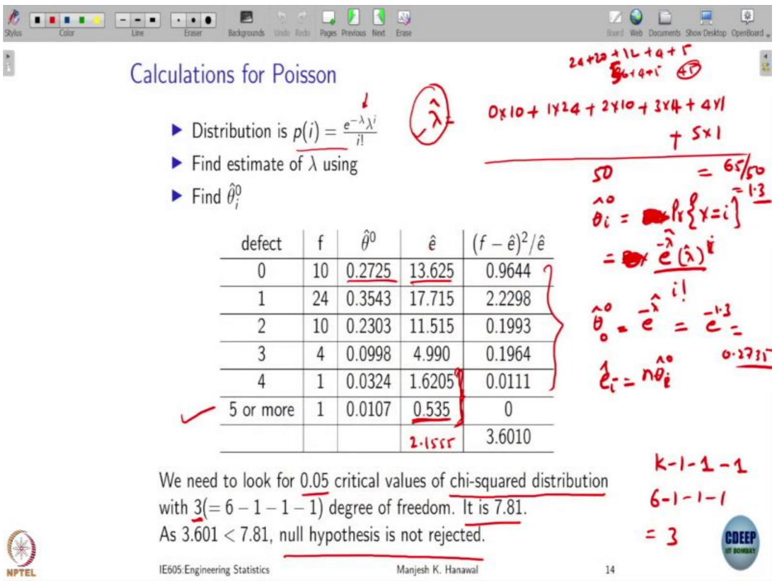
And now, I know that n theta i is exactly equals to the ei, ei minus fi square divided by fi and this too i equals to 1 to k fi epsilon square. And now, if you recall, fi has the term theta i 0 divided by fi by n, it has the term maybe raised to the power n, and many terms like has terms for n greater than has many terms like this, terms and now for. So, what was that like we have these terms like this, which is it was like j greater than or equals to 3.

And we notice that as n tends to infinity, by law of large number we know that fi by n is going to converge to 0. So, because of that, all when n tends to infinity, this epsilon i is going to converge to 0. And that is why we can say that as n tends to infinity minus 2 log of t is approximately or in fact is equals to summation ei minus fi square by ei which is exactly our Q matrix.

And since we already argued that minus 2 log T is Q distributed is Chi-square distributed with k minus stream, we can say that, we can approximate Q also to have Chai-squared k minus 1 degrees of freedom. So, I hope so, there is a small difference here the way we wrote this Q here, where our denominator is v i, but what we got on here is in the denominator as fi. So, please check this thing and make sure that whether what we actually wanted Q here to be one, the one with the denominator, fi, or I made any calculation mistakes in this please do verify that, fine.

(Refer Slide Time: 19:35)





Now, how to apply this? We now just saw that whenever I have a data and I need to do Q test, all I need to do was compute my statistic Q, and check whether it is greater than or equals to z alpha to get a alpha significance test. To quickly do that, let us look into an example. Let us say that our quality control engineering has taken 50 samples of size 13.

So, a quality engineering is gathering, let us say this is sample 1, where he is taking 13 samples from this is like 1 batch. And similarly, he take another lot of 13 samples from the production process like that he does 50 batches. And out of in each of these batches, he looks, the number of defects for the samples. And the table gives the number of defects that he has observed in how many defects here are going to observe in this samples of batch size 13.

And we want to test our null hypothesis, at level .05, whether the defects are going to follow either Poisson, binomial distribution, I am going to do the exercise computation for the binomial distribution. And you can try out similarly for the binomial. So, notice that we have only given Poisson distribution, and we do not know the value of lambda, we are only given this. And what you are told is in a 0 defect were found in 10 of this batches. Only one defect was found in 24 of them, 2 defects are found in 10. Like that, and 6 or more defects were not found in any of the samples. Now, notice that the number of classes we have is only this now 0, 1, 2, 3, 4, 5, 6. That is my classes are actually 7 here. Starting from 0, 1, like this.

Now, for the Poisson distribution, I know it has the shape. This for its probability mass function, here, my lambda is not known. But I can estimate lambda from the samples. So, how can I estimate lambda from the samples, I have this value that the expected value, I know 0 defects were found in 10. And one defect in 24, 2 in 10, 3 in 4, 4 in 1, and 5 in 1.

So, notice that even though I said k equals to 7 here, but in none of the samples, I found any, like 6 or more defects. So, that means basically the number of possible values of the number of defects that are being taken as 0, 1, 2, 3, 4, 5, 6. Because of that, my actual distinct classes are only 5, or like 6 here, including the value of 0, I get my total number of 6 because I am never observing 6 or more defects. So, I do not need to take that as one of the class.

Now, I have this value divided by total number of samples is 50. If you compute this value, you are going to get the value of lambda hat, whatever that lambda hat you can find out and now how you are going to compute theta i hat 0, the theta i hat 0, you are going to compute as we are going to compute as 50 into probability that x i equals to 0 which is 50 into e to the power minus lambda hat, lambda hat into i, sorry, i here, divided by i factorial, whatever that comes.

So, that is the value. Sorry, this is theta hat is simply the probability, so you can calculate this and that is what we have done here. And you see that for theta i, when i 0 this is like e to the power minus lambda hat. It is simply and I think whatever these computes, I think here, you can quickly compute 24 plus, 20 plus, 12 plus, 4, plus 5 that is 24, 36, 4 plus 5.

Whatever the value of lambda hat you get, you can plug in. And I just verify that indeed I calculated it correctly here once you plug in this value, we are going to get this theta hat values and ei values is nothing but n into theta i values here, just I want to revisit this, is observed 10, and 1 into 24, just let me see if I have this here, 24, 20, 12, 4 and 5. 24, this is 40, 44, 56, this is 56, plus 4 and 5, 50.

So, you will end up this value to be close to total 65 by 50, which is like 1.3. And in fact, you can see that minus 1.3 is point, if you see that this is exactly equals to 0.2735. And based on this computation, you can compute what is Heap bar here. And similarly, you can compute all of this.

And I noticed that when you do this for this class 5, you are ending with a value 0.5, which is less than 1. And the frequency cannot be, the count here, basically, we are counting the number of expected frequency of a class which cannot be less than 1. So, we may want to merge these 2 things, and this is going to give us a value of 2.1 What is this 5, 5, 5 am I correct 5, 5, 5, and then do the recomputation.

So basically, we further restricted my I did one more grouping here. So, now I can only compute basically I kind of ignoring this class as well now, and now based on that I can compute as required f minus e bar and I am going to get this value, and then I am going to sum that that is going to be my value. Now, if you want to check this critical value at 0.05, I need to first decide what is the threshold I should set.

In this case, it so happened that we need to take the critical value from Chi-square distribution with 3 degrees of freedom. And I will just comment about why this is 3 here, and this value happens to be 7.81. And whatever the value we have obtained Q value 0.360 here is going to be less than 7.81. So, in this case, that is why we are not going to reject this null hypothesis as per our criteria.

Now, the question is why we take Chi-square distribution with 3 degrees of freedom. This is because, first of all, we started with k equals to 6 here. And we need to take the degrees of freedom to k minus 1, but we further reduce the class here because my e hat happens to be less than 1 for my class 5 or more. So, I needed to reduce my class 1 further, and I also had an estimation for this lambda and that will also reduce my thing 1 that is why the 6 minus 1, minus 1, minus 1 which is 3, which I am taken as my degrees of freedom.

I hope this clarifies. And similarly, you can compute this test for the binomial distribution also, you will see that even the binomial test for the binomial also said that your null hypothesis is not rejected. So, both our tests say that both of these cannot be rejected. And obviously, both cannot be correct. So, it is in this case, we can conclude that our samples are not enough to say conclusively which one is possibly to be rejected. So, this is all about how to use your Chi-square distribution. Now, we will talk about the other test.