

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

**Lecture – 59**  
**Test for Goodness of Fit – I**

So far in this course on Statistical Inference, the portion for which I have covered the testing of hypothesis, I have considered two major approaches. One is the theory of most powerful test, uniformly most powerful test or uniformly most powerful unbiased test etcetera. The theory which has been developed following Neyman-Pearson work and we derived the exact test for several situations, especially for the parameters of the normal distributions.

The another approach, which I considered was the consideration of likelihood ratio test, and there also we were able to derive the exact tests for various situations. One important point which I mentioned in the likelihood ratio test is that sometimes when the exact distribution of the test statistic in the likelihood ratio test is not possible, the asymptotic distribution of minus twice log of lambda, where lambda is the likelihood ratio is a chi square distribution under the null hypothesis, and that is helpful.

Now, today we will discuss another type of tests. Many times we are saying, we want to test whether the data comes from a particular distribution that means whether the data is from binomial distribution, whether the data is from a Poisson distribution, whether the data is from an exponential distribution etcetera. And then how to test that? So, a class of tests for this has been developed, they are called goodness of fit test.

(Refer Slide Time: 02:07)

Lecture 36

Chi-Square Test for Goodness of Fit

Suppose we are having a random sample from a population  $F(x)$  (may depend upon a parameter  $\theta$ )

We want to test

$$H_0: F(x) = F_0(x) \quad \forall x$$
$$H_1: \quad \neq \quad \text{for at least some } x.$$

Case I:  $F_0(x)$  is completely specified.

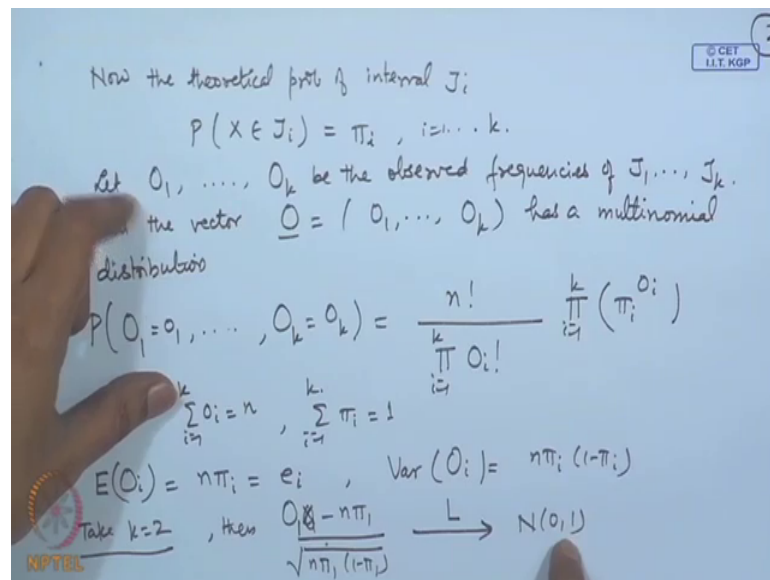
We divide the range of the dist<sup>n</sup>.  $F_0(x)$  into  $k$  mutually exclusive and exhaustive intervals say  $J_1, \dots, J_k$

However, I will restrict attention to only chi square test for this situation, so the model is like this. Suppose, we are having a random sample from a population say  $F(x)$ ; now, this  $F(x)$  may depend upon a parameter, it may not depend upon a parameter. So, may depend upon a parameter  $\theta$ . In the nonparametric situations, we may not consider this  $\theta$  here.

Now, we want to test hypothesis say  $F(x)$  is equal to some known distribution  $F_0(x)$  for all  $x$ , against not equal, so that means for at least some  $x$ . Now, there can be two cases. I am considering case I,  $F_0(x)$  is completely specified. Now, we want to test whether the data  $x_1, x_2, \dots, x_n$  is actually from this distribution  $F_0(x)$ .

So, the method of applying this chi square test for goodness of fit is the following, we classify or you can say we divide the range of the distribution  $F_0(x)$  into  $k$  mutually exclusive and exhaustive intervals. Let us say these intervals say  $J_1, J_2, \dots, J_k$  that means, what we are saying is that this  $J_i$ 's are disjoint. And the union of  $J_i$ 's is the range of the variables that is the range of the distribution, suppose it is from 0 to infinity or it is from minus infinity to infinity or if it is a finite interval from  $a$  to  $b$ .

(Refer Slide Time: 04:48)



Now, the theoretical probability of interval  $J_i$  that is probability of  $X$  belonging to  $J_i$  is equal to say  $\pi_i$ ,  $i$  is equal to 1 to  $k$ . Let us consider say, so when we are considering the observations divided into intervals  $J_1, J_2, J_k$ , then there will be observed frequencies. So, let  $O_1, O_2, O_k$  be the observed frequencies of  $J_1, J_2, J_k$  ok.

Then the vector  $O$  of this observed frequencies  $O_1, O_2, O_k$  has a multinomial distribution. We can write it like this probability of say  $O_1$  is equal to small  $o_1$  and so on,  $O_k$  is equal to small  $o_k$  that is equal to  $n$  factorial divided by product of  $O_i$  factorial,  $i$  is equal to 1 to  $k$ , then product of  $\pi_i$  to the power  $O_i$ ,  $i$  is equal to 1 to  $k$ , where sigma of  $O_i$  is equal to  $n$ , and sigma of  $\pi_i$  is equal to 1.

Now, expectation of say  $O_i$  that will be equal to  $n \pi_i$ , let us call it say  $e_i$ . And variance of  $O_i$  that will be equal to  $n \pi_i (1 - \pi_i)$ . In the multinomial distribution, these are the statements that are satisfied. Now, let us take say  $k$  equal to 2 that means, only two classifications are there. Then actually it becomes binomial distribution.

And for binomial distribution, we have the statement  $X - n \pi_1$  divided by square root  $n \pi_1 (1 - \pi_1)$ . So, this  $X - n \pi_1$  is  $O_1$ , this converges in distribution to normal  $0, 1$ . This is this notation for convergence in distribution. The asymptotic distribution of  $O_1 - n \pi_1$  divided by square root  $n \pi_1 (1 - \pi_1)$  is a standard normal, this is by the binomial approximation to the normal distribution.

(Refer Slide Time: 08:03)

$$\Rightarrow \frac{(O_1 - n\pi_1)^2}{n\pi_1(1-\pi_1)} \xrightarrow{L} \chi_1^2$$

For  $k=2$ ,  $O_2 = n - O_1$ , we can then show that

$$\frac{(O_1 - n\pi_1)^2}{n\pi_1} + \frac{(O_2 - n\pi_2)^2}{n\pi_2} = \frac{(O_1 - n\pi_1)^2}{n\pi_1(1-\pi_1)}$$

i.e.  $\sum_{i=1}^2 \frac{(O_i - n\pi_i)^2}{n\pi_i} \xrightarrow{L} \chi_1^2$

For general  $k$ ,  $T = \sum_{i=1}^k \frac{(O_i - n\pi_i)^2}{n\pi_i} \xrightarrow{L} \chi_{k-1}^2$ .

We calculate  $e_i = n\pi_i$  from the known dist<sup>n</sup>  $F_0(x)$  (under  $H_0$ ).

Now, this implies that if I consider  $O_1 - n\pi_1$  square divided by  $n\pi_1(1-\pi_1)$ , then the asymptotic distribution of that will be chi square on 1 degree of freedom. Now, for  $k$  equal to 2, we will have  $O_2$  is equal to  $n - O_1$ . So, we can then show that  $(O_1 - n\pi_1)^2 / n\pi_1 + (O_2 - n\pi_2)^2 / n\pi_2$ , that is equal to  $(O_1 - n\pi_1)^2 / n\pi_1(1-\pi_1)$  that is the distribution of  $(O_i - n\pi_i)^2 / n\pi_i$ ,  $i$  is equal to 1 to 2 that is asymptotically chi square 1.

For general  $k$ , we can proceed in the same way. We can show that the distribution of  $(O_i - n\pi_i)^2 / n\pi_i$ ,  $i$  is equal to 1 to  $k$  is asymptotically chi square on  $k - 1$  degrees of freedom. Now, when so this  $\pi_i$ 's are the theoretical probabilities. We calculate this  $e_i$ 's that is equal to  $n\pi_i$  from the known distribution  $F_{naught}(x)$  that is under  $H_{naught}$ .

(Refer Slide Time: 10:40)

The image shows a handwritten derivation of the chi-square test statistic. It starts with the formula  $T = \sum_{i=1}^k \frac{(O_i - e_i)^2}{e_i}$ , which is asymptotically equivalent to  $\chi^2_{k-1}$ . It then states that we can set up a test for goodness of fit by rejecting the null hypothesis if  $T > \chi^2_{k-1, \alpha}$ . The derivation then simplifies the formula to  $T = \sum \frac{O_i^2 + e_i^2 - 2e_i O_i}{e_i}$ , which is further simplified to  $\sum \frac{O_i^2}{e_i} + \sum e_i - 2 \sum O_i$ . A box notes that  $\sum O_i = n$  and  $\sum e_i = n$ , where  $n$  is the total number of observations. The final simplified formula is  $T = \sum \frac{O_i^2}{e_i} + n - 2n = \sum \frac{O_i^2}{e_i} - n$ .

So, this expression then becomes, then this T is equal to sigma O i minus e i square divided by e i, i is equal to 1 to k, this is asymptotically chi square on k minus 1 degrees of freedom. So, we can set up a test for goodness of fit by reject H naught, if T is greater than chi square k minus 1 alpha, where chi square k minus 1 alpha is the upper 100 alpha percent point of the chi square k minus 1 distribution.


Now, here you have some sort of simplification also available. In place of this, see this we can write sigma O i square plus e i square minus twice e i divided by e i that is equal to sigma O i square by e i plus sigma e i minus twice n twice O i e i, so that is equal to minus twice O i sigma. Now, sigma O i is n, sigma e i is n that is the total number total number of observations, so that is equal to sigma O i square by e i plus n minus twice n that is equal to sigma O i square by e i minus n. So, this is a simplified formula for T here.

(Refer Slide Time: 12:53)

Example: A random variable simulator is used to generate 1000 values for a  $U[0,1]$  r.v. The values are classified into intervals  $[0, 0.1)$ ,  $[0.1, 0.2)$ , ...,  $[0.9, 1]$ . The observed frequency dist is as follows

	0-0.1	0.1-0.2	0.2-0.3	0.3-0.4	0.4-0.5	0.5-0.6	0.6-0.7	0.7-0.8	0.8-0.9	0.9-1
O <sub>i</sub>	112	101	94	99	108	93	94	100	104	95

© CET I.I.T. KGP 5



Let me give an example here. A random variable simulator is used to generate 1000 values for a uniform 0, 1 random variable. The values are classified into intervals like 0 to 0.1, 0.1 to 0.2 and so on, 0.9 to 1 ok. The observed frequency distribution is as follows.

So, here I will present interval and here the number of terms. So, these are O<sub>i</sub>'s between 0 to 0.1 between 0.1 to 0.2 between 0.2 to 0.3, so there are 112 values here, between 0.1 to 0.2 it is 101, 94, 0.3 to 0.4 99, 0.4 to 0.5 108, 0.5 to 0.6 93, 0.6 to 0.7 94, 0.7 to 0.8 100, 0.8 to 0.9 104, and 0.9 to 1, there are say 95 values.


(Refer Slide Time: 15:49)

into intervals  $[0, 0.1)$ ,  $[0.1, 0.2)$ , ...,  $[0.9, 1]$ . The observed frequency dist<sup>n</sup> is as follows

	0-0.1	0.1-0.2	0.2-0.3	0.3-0.4	0.4-0.5	0.5-0.6	0.6-0.7	0.7-0.8	0.8-0.9	0.9-1
$O_i$	112	101	94	95	108	93	94	100	104	95
$e_i$	100	100	100	100	100	100	100	100	100	100

We want to test whether the simulator is working properly.  
 $H_0: F(x) = U(x) \neq x$  where  $U(x)$  is cdf of  $U[0, 1]$  dist<sup>n</sup>  
 $H_1: \text{not}$

Under  $H_0$   $P([0, 0.1]) = 0.1, \dots, P([0.9, 1]) = 0.1$   
 $\pi_i = 0.1, i=1, \dots, 10, k=10$   
 $e_i = n\pi_i = 1000 \times 0.1 = 100$

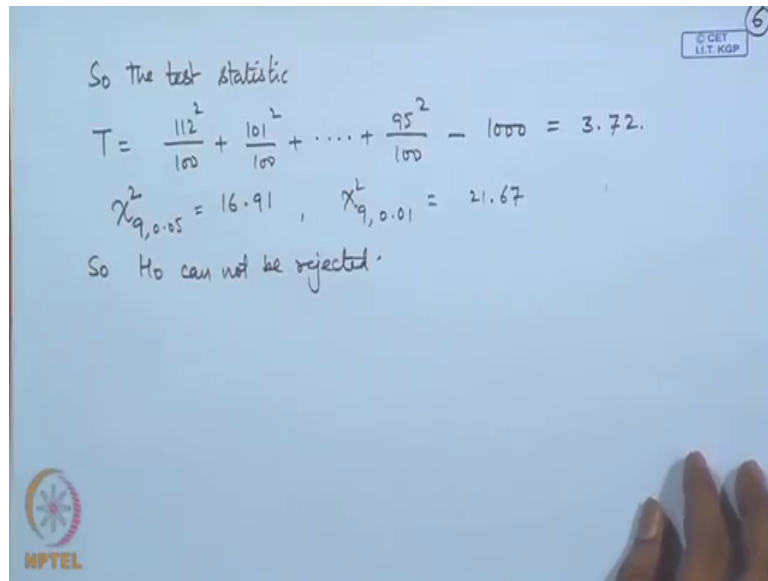


So, we want to test, whether the simulator is working properly. So, what does it mean that the simulator is working properly, if it is working properly, then the observations that we have got, and we have classified like this should fit into a uniform distribution on the interval 0 to 1 that means, you want to test the hypothesis, whether  $F(x)$  is equal to say  $U(x)$ , where  $U(x)$  is cdf of uniform 0, 1 distribution. Against the hypothesis not, that  $F(x)$  is not  $U(x)$ .

So, in order to apply the chi square test for goodness of fit, I need to calculate the probabilities of the intervals. Now, if I am assuming under  $H_0$  probability of this each interval of this nature that is 0 to 0.1, this will be 0.1 and so on, probability of each interval 0.9 to 1 that will also be equal to 0.1 that is these are my  $\pi_i$ 's, so  $\pi_i$ 's are 0.1 for  $i$  is equal to 1 to 10 that is  $k$  is equal to 10, we are having 10 classes here.

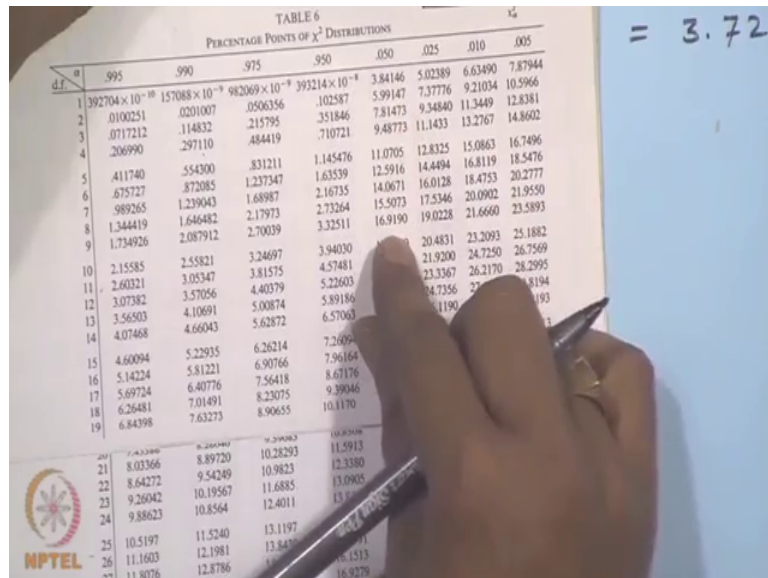
Now, based on this  $e_i$ ,  $e_i$  is equal to  $n\pi_i$  that is equal to 1000 times 0.1 that is equal to 100; that means for each class in this case  $e_i$  is only 100. So, if we want to apply this chi square test, I need to calculate  $\sum (O_i - e_i)^2 / e_i$  or I can also calculate  $\sum O_i^2 / e_i - n$ . So, these things are not difficult now.

(Refer Slide Time: 18:41)



This T statistic then, so the test statistic T that is equal to 112 square by 100 minus 101 square by 100 plus all the terms will be plus 95 square by 100 minus 1000. Now, this can be easily evaluated, this value turns out to be 3.72 approximately. Now, we need to look at the value of chi square on 9 degrees of freedom at alpha level of significance.

(Refer Slide Time: 19:36)



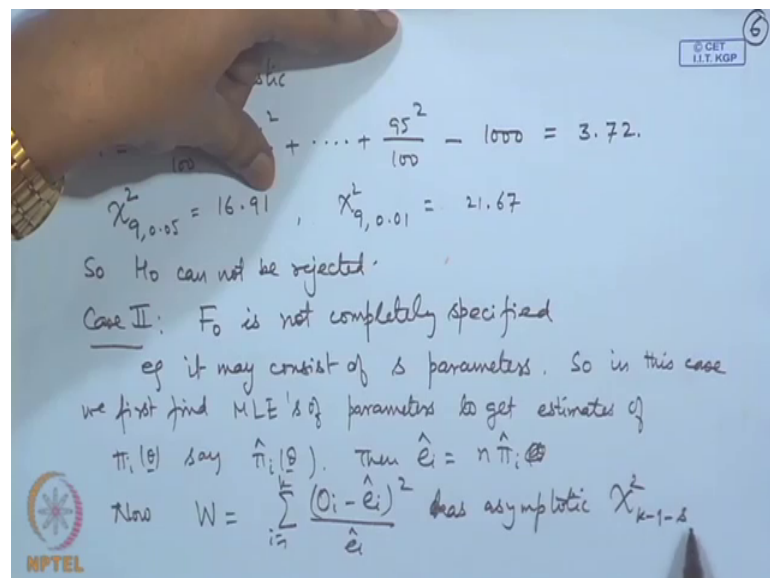
Now, you can see the tables of chi square distribution, so I will show from here. So, tables of chi square distribution, suppose I look at say 0.05, the value is equal to 16.91.



So, chi square 9 on 0.05 is say 16.91, we may also look at say chi square 9, 0.01 that is equal to 21.67 and so on. So, these values you can see they are larger.

So,  $H_0$  cannot be rejected that means, this simulator seems to be working properly. Now, in this particular case, when I wanted to test whether  $F$  is equal to  $F_0$ , then  $F_0$  was completely specified, because I said uniform 0, 1. But, there can be cases, when the distribution maybe known, but it may not be completely known. For example, it may depend upon a parameter, suppose I say uniform 0 theta or I say binomial  $n, p$ , where  $p$  is not known. So, I want to test whether the data is from a binomial distribution, whether it is from a normal distribution, but the parameters may not be specified. In that case, we need to actually estimate those parameters from the data.

(Refer Slide Time: 20:57)

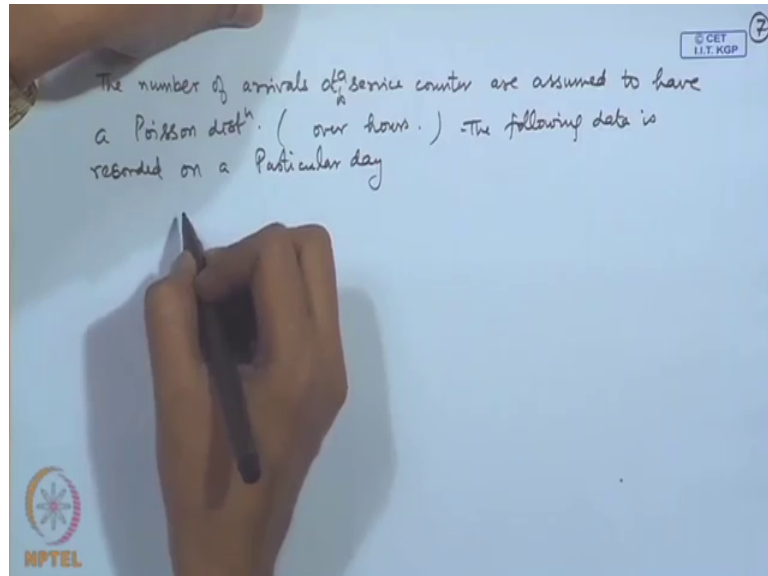


So, I will consider case 2,  $F_0$  is not completely specified. For example, it may consist of say  $s$  parameters. If it consists of parameters, then these probabilities need to be estimated, so what we do? So, in this case, we first find say maximum likelihood estimators of parameters to get estimates of  $\pi_i(\theta)$ , say  $\hat{\pi}_i(\theta)$  ok. Then  $\hat{e}_i$  that is equal to  $n$  times  $\hat{\pi}_i(\theta)$ , so I am just writing  $n \hat{\pi}_i(\theta)$ .

Now, let me say  $W$  that is equal to  $\sum_{i=1}^k \frac{(O_i - \hat{e}_i)^2}{\hat{e}_i}$  has asymptotic chi square distribution on  $k - 1 - s$  degrees of freedom. So, in place of  $k - 1$  now it is becoming  $k - 1 - s$ , because this is

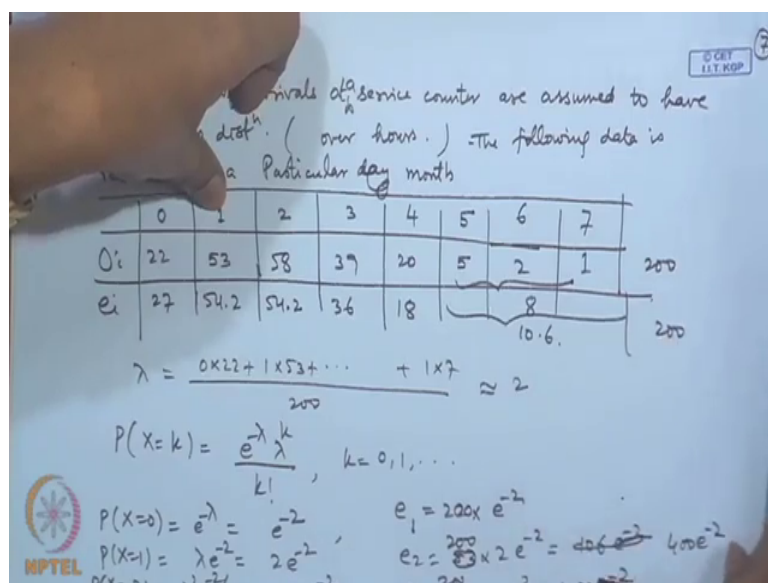
the number of the degrees of freedom which are now reduced because of allocation or you can say estimation of the parameters which were unknown.

(Refer Slide Time: 23:05)



Let me give one example of the situation here. So, we have the data say on certain arrivals ok, the number of arrivals at a service counter assumed to have a Poisson distribution ok, the time period is over hours ok. Now, the following data is recorded on a particular day.

(Refer Slide Time: 24:19)



So, the no arrivals were observed 22 times on a particular month let me say sorry in place of day. And 1 arrival observed on 53 times, 2 arrivals were observed 58 times, 3 arrivals were observed 39 times, 4 arrivals were observed 20 times, 5 arrivals were observed 5 times, and 6 arrivals were observed 2 times, 7 arrivals were observed 1 time. So, the total data was 200.

Now, when we apply the chi square test, the expected frequency for each class should be greater than or equal to 5, if it is below that, then we need to combine. Now, here you can see the observed frequency for these two classes is less than 5. Naturally, when we consider the estimation, it may turn out to be less. So, what we do, we merge this. We merge this into one group say 8, and now we do not know the parameter of the Poisson distribution.

So, we estimate lambda by the mean, so we consider 0 into 22 plus 1 into 53 and so on plus 1 into 7 divided by 200 that is approximately 2; so, then expected frequencies are calculated using the probability distribution. So, we have  $e$  to the power minus lambda, lambda to the power  $k$  by  $k$  factorial,  $k$  equal to 0, 1 to and so on.

So, what is the probability  $X$  equal to 0? That is  $e$  to the power minus lambda that is equal to  $e$  to the power minus 2. So, this is  $O_i$ , this is  $e_i$ . So,  $e$  corresponding to the first grouping that is equal to 22 into  $e$  to the power minus 2. So, this can be calculated. Then we are having probability  $X$  equal to 1 for example that is equal to lambda  $e$  to the power minus lambda that is twice  $e$  to the power minus 2. So,  $e$  to then will become 53 into 2  $e$  to the power minus 2 that is 106  $e$  to the power minus 2.

(Refer Slide Time: 27:23)

Handwritten notes on a whiteboard showing a table of observed frequencies and the calculation of probabilities for a Poisson distribution with  $\lambda = 2$ .

0	1	2	3	4	5	6	7	8	...	200
		58	39	20	5	2	1			

$$= \frac{0 \times 22 + 1 \times 53 + \dots + 1 \times 7}{200} \approx 2$$

$$P(X=k) = \frac{e^{-\lambda} \lambda^k}{k!}, \quad k=0,1,\dots$$

$$P(X=0) = e^{-\lambda} = e^{-2}$$

$$P(X=1) = \lambda e^{-\lambda} = 2e^{-2}$$

$$P(X=2) = \frac{\lambda^2 e^{-\lambda}}{2!} = 2e^{-2}$$

$$e_1 = 22 \times e^{-2}$$

$$e_2 = 53 \times 2 e^{-2} = 106 e^{-2}$$

$$e_3 = 58 \times 2 e^{-2}$$

MPTEL logo is visible in the bottom left corner.

Similarly, you can calculate probability X equal to 2 say that is equal to lambda square e to the power minus 2 by 2 factorial that is equal to again 2 e to the power minus 2, so e 3 that will be equal to 58 into 2 e to the power minus 2.

(Refer Slide Time: 27:45)

Handwritten notes on a whiteboard showing a chi-square test for goodness of fit to a Poisson distribution.

$$P(X=3) = \dots$$

$$\sum \frac{(O_i - e_i)^2}{e_i} = 2.33$$

$$\chi^2_{5, 0.05} = 11.1433$$

$$H_0 \text{ i.e. hyp that the data is from Poisson dist}^n \text{ can not be rejected}$$

0, 1, 2, 3, 4  
 $\geq 5$   
 $k=5$

© CET I.I.T. KGP logo is visible in the top right corner.

MPTEL logo is visible in the bottom left corner.

Similarly, you can calculate say probability X is equal to 3 and so on. And I will combine these three probabilities that is you can consider them as probability X greater than or equal to 5. And this values will then turn out to be, sorry this is slightly mistaken here, this will be 200 here, you are multiplied by n, so this is 400 e to the power minus 2, this

is also  $400 e^{-2}$ . So, now let me write down the values here. These values are 27, 54.2, 54.2, 36, 18, and this last three we add up it is 10.6, this total is again 200 here.

Now, we calculate this  $\sum (O_i - e_i)^2 / e_i$ , this can be calculated it turns out to be 2.33. Now, let us compare from the chi square value on how many groups we have made, we have made five groups that is 0, 1, 2, 3, 4 and greater than or equal to 5. So, there are six groups here. So, we look at the chi square value on 5 degrees of freedom, and this value then turns out to be 11 point say 0.05, it is 11.143. So, once again we can say that  $H_0$  that is the hypothesis that the data is from Poisson distribution cannot be rejected.

Now, this chi square test for goodness of fit has some further generalizations or you can say further applications. Let me give the application of it for testing for independence in a contingency table. Earlier you remember that when we were discussing the theory of UMP unbiased tests, I considered a 2 by 2 contingency table, and I actually obtained an exact UMP unbiased test.

However, that UMP unbiased test dependent upon was depending upon certain discrete distribution. And so it was not very convenient to apply, you need to look at the tables and moreover it is a discrete distribution, therefore the randomization will be required to get the exact level of significance there.