Probability and Statistics

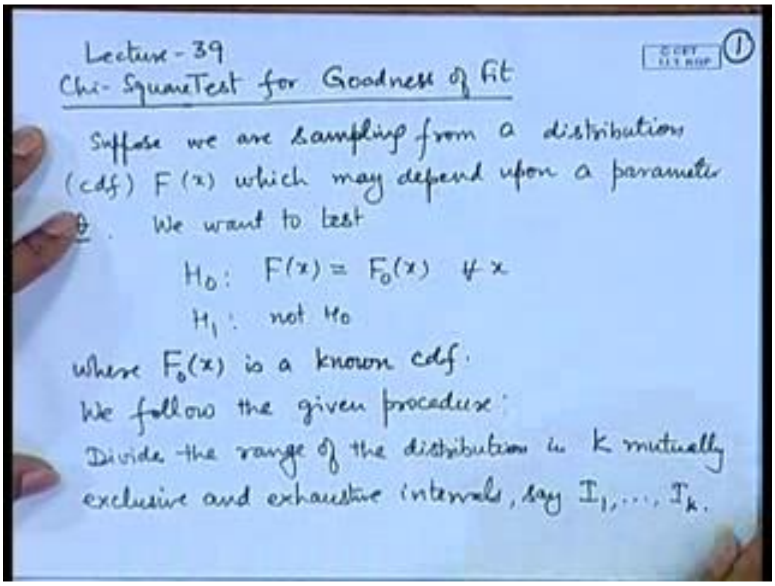
Prof. Somesh Kumar
Department of Mathematics
Indian Institute of Technology, Kharagpur

Lecture - 76
Chi-Square Test for Goodness Fit – I

In the situations so far for testing of hypothesis problems, we have considered that the random sample comes from a certain population. So, we have assumed the form of population to be say normal or say exponential and then we want to test about the parameters of that population. For example, we want to test about the mean of a normal distribution, we want to test about the variance of a normal distribution or we want to test about the scale parameter of an exponential distribution. So, here the situation is that we assume that the form of the distribution is known to us, only the parameters of the population are not known.

However there are other situations where we have a data and we want to know that from which type of population that data has come and so that mean we may be liking to estimate the population; that means, the distribution or we want to test about the distribution. So, here we will talk about the testing that a particular distribution is say capital F is equal to some f naught. So, for this situation an approximate test as been proposed which is called chi square test for goodness of fit.

(Refer Slide Time: 01:52)



Lecture - 39

Chi- SquareTest for Goodness of Fit

Suppose we are sampling from a distribution (cdf) $F(x)$ which may depend upon a parameter θ. We want to test

$$H_0: F(x) = F_0(x) \quad \forall x$$
$$H_1: \text{not } H_0$$

where $F_0(x)$ is a known cdf.

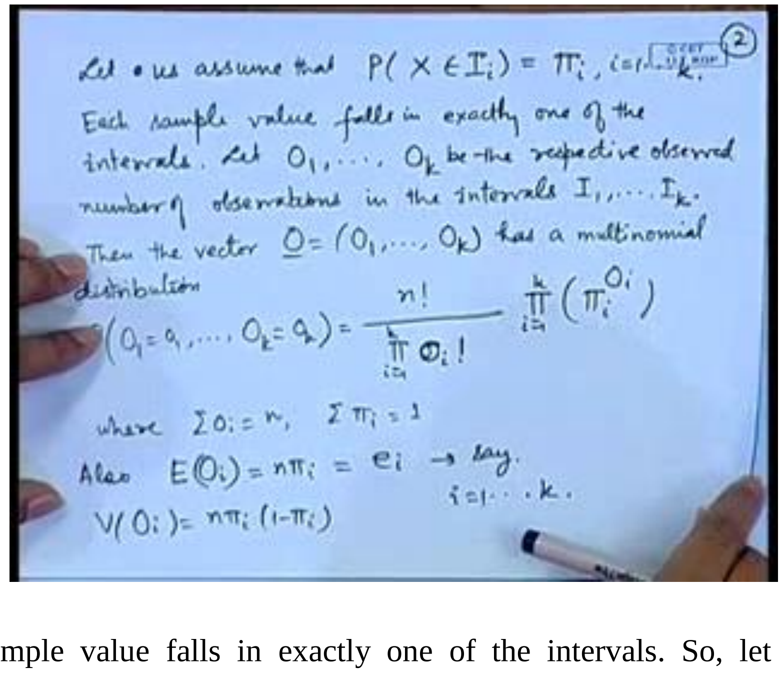
We follow the given procedure:

Divide the range of the distribution in k mutually exclusive and exhaustive intervals, say $I_1, \dots, I_k$.

Let us have the situation of the form suppose, we are sampling from a distribution. So, let me write distribution function cdf say F x. Of course, there may be a situation where it depends upon certain parameters. So, we may write that thing which may depend upon a parameter say theta. So, this theta could be having several components also. So, we want to test say H naught F x is equal to F naught x for all x against H 1 not H naught; that means, this is not true for some points, at least where F naught is a known cdf; that means, you want to test whether the data which has been collected comes from a given distribution F naught x.

In the chi square test for goodness of fit the procedure is as follows, we follow the given procedure. So, divide the range of the distribution in k mutually exclusive and exhaustive intervals. Let us call them intervals as I 1, I 2, I k. So, now, each value will fall exactly in 1 of the intervals because they are mutually exclusive and exhaustive. Let us also assume that let us assume that probability of X being in an interval I is pi i for i is equal to 1 to k.

(Refer Slide Time: 04:24)



2

Let us assume that $P(X \in I_i) = \pi_i$, $i = 1 \ldots k$.
Each sample value falls in exactly one of the intervals. Let $O_1, \ldots, O_k$ be the respective observed number of observations in the intervals $I_1, \ldots, I_k$.
Then the vector $\underline{O} = (O_1, \ldots, O_k)$ has a multinomial distribution

$$(O_1 = o_1, \ldots, O_k = o_k) = \frac{n!}{\prod_{i=1}^{k} O_i!} \prod_{i=1}^{k} (\pi_i^{O_i})$$

where $\sum O_i = n$, $\sum \pi_i = 1$

Also $E(O_i) = n\pi_i = e_i \rightarrow$ say.
$i = 1 \ldots k$.

$V(O_i) = n\pi_i(1-\pi_i)$

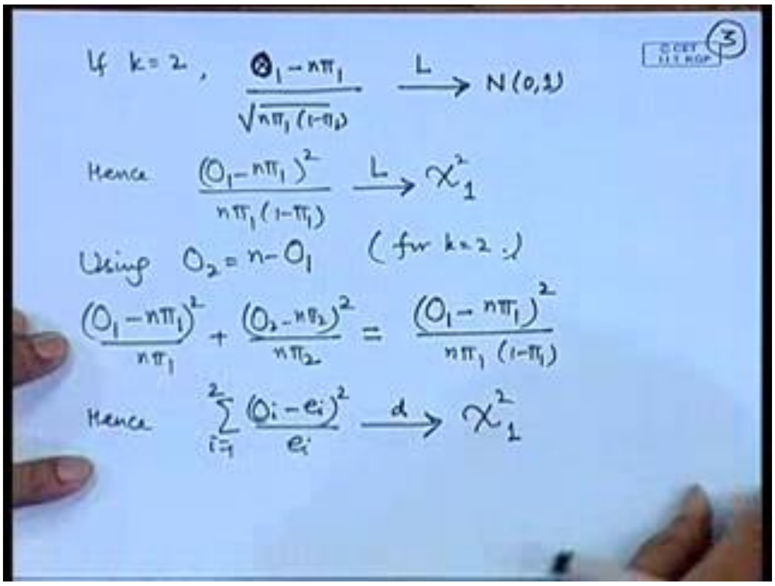
Now each sample value falls in exactly one of the intervals. So, let us define the observed frequencies, let O 1, O 2, O k be the respective observed number of observations in the intervals I 1, I 2, I k. So, what we are observing? We are observing x 1, x 2, x n, now you see some of these x i's will belong to interval I 1, some of the x i's will belong to interval I 2 and so on. So, we make this break up.

Let O 1 denote the number of x i’s in the interval I 1, let O 2 denote the number of x i’s in I 2, let O k denote the number of x i’s in I k then O 1, O 2, O k, these are called observed frequencies of the data. So, now, you see here you have k categories unlike binomial distribution where you have 2 categories here you have k categories. So, if I find out the distribution of O 1, O 2, O k, this will become a multinomial distribution. So, we will write it in this form then the vector O that is O 1, O 2, O k has a multinomial distribution n factorial divided by product of O i factorial then product of pi i to the power O i i is equal to 1 to k.

Where sigma of O i is n sigma of pi i is 1 that is pi i is the probability of the i th interval therefore, probability of O i is equal to a small o i that will be given by this multinomial function, also we will have from the properties of the multinomial distribution expectation of O i, this will become equal to n pi, I we call it e i and variance of O i that will become equal to n pi i into 1 minus pi i for i is equal to 1 to k.

Let us take the case of 2 categories then let us see how the test can be conducted.

(Refer Slide Time: 08:21)



If $k=2$, $\dfrac{O_1 - n\pi_1}{\sqrt{n\pi_1(1-\pi_1)}} \xrightarrow{L} N(0,1)$

Hence $\dfrac{(O_1 - n\pi_1)^2}{n\pi_1(1-\pi_1)} \xrightarrow{L} \chi^2_1$

Using $O_2 = n - O_1$ (for $k=2$:)

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

Hence $\displaystyle\sum_{i=1}^{2} \frac{(O_i - e_i)^2}{e_i} \xrightarrow{d} \chi^2_1$

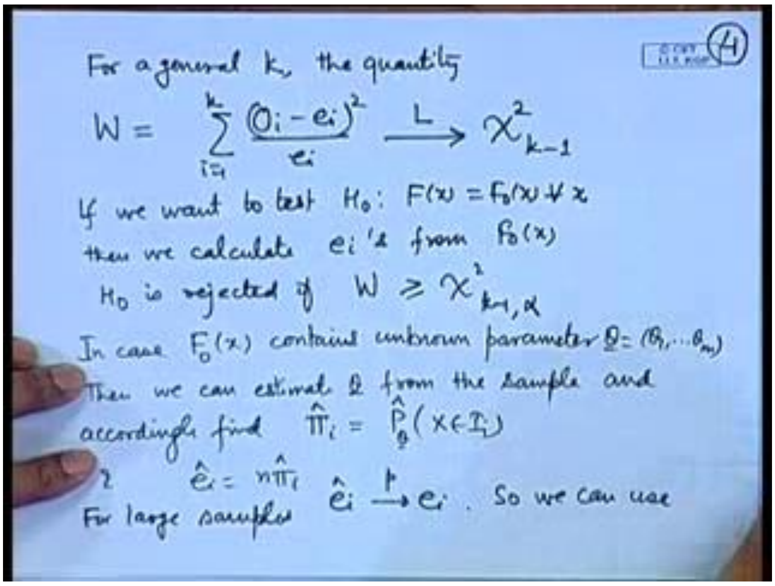
If k is equal to 2 then if I look at x 1 minus n pi 1, in place of this, I will write O 1. So, O 1 minus n pi 1 divided by root n pi 1 into 1 minus pi 1, this is converging in distribution to normal 0 1, this is from the property of the binomial distribution that the distribution of x minus n pi divided by root n p cube is asymptotically normal that is the normal approximation to the binomial distribution.

If we utilize that because for k equal to 2, this multinomial has become binomial therefore, the distribution of O 1 is binomial n pi 1. So, O 1 minus n pi 1 divided by square root n pi 1 into 1 minus pi 1 is asymptotically normal distribution. So, if I take the square of this then this is asymptomatically chi square distribution on 1 degree of freedom because a square of a normal distribution standard normal variable is a chi square variable.

Now, O 2 is n minus O 1, this is for the case k equal to 2. So, you can easily that that if I write down O 1 minus n pi 1 square by n pi 1 plus O 2 minus n pi 2 square by n pi 2 then so here you substitute O 2 is equal to n minus O 1 and pi 2 is equal to 1 minus pi 1 then after simplification because this will become 1 minus pi 1, we can take LCM and adjust the terms this will give simply O 1 minus n pi 1 square divided by n pi 1 into 1 minus pi 1.

What we are observing that sigma O i and this time I called e i square by e i, i is equal to 1 to this is having asymptomatically chi square distribution on 1 degree of freedom.

(Refer Slide Time: 11:03)



4

For a general k, the quantity

$$W = \sum_{i=1}^{k} \frac{(O_i - e_i)^2}{e_i} \xrightarrow{L} \chi^2_{k-1}$$

If we want to test $H_0: F(x) = F_0(x) \; \forall x$

then we calculate e_i's from $F_0(x)$

H_0 is rejected if $W \geq \chi^2_{k-1, \alpha}$

In case $F_0(x)$ contains unknown parameter $\underline{\theta} = (\theta_1, \dots \theta_m)$

Then we can estimate $\underline{\theta}$ from the sample and accordingly find $\hat{\pi}_i = \hat{P}_{\theta}(X \in I_i)$

$\hat{e}_i = n\hat{\pi}_i$

For large samples $\hat{e}_i \xrightarrow{P} e_i$. So we can use

If we generalize this thing in place of 2 if i write for a general k, the quantity let me call it W that is equal to sigma O i minus e square by e i i is equal to 1 to k, this is having asymptomatically chi square distribution on k minus 1 degrees of freedom I am using L and d as for the asymptomatic distribution here.

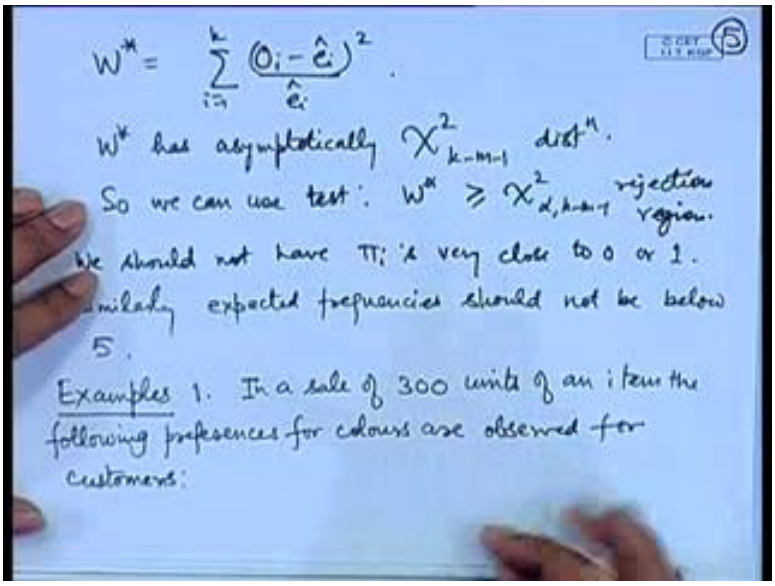
If we want to test the hypothesis, if we want to test H naught F x is equal to F naught x then we calculate e i's from F naught x that is under the distribution F naught, what is the probability of the i th interval. So, that is pi i and if I multiply by n i will get e i so, but of course, this is calculated from the known distribution this one here. So, you can then consider the difference between the observed frequency and the expected frequency squared and divided by the expected frequency.

You can see that if this hypothesis is true then the differences between O i's and e i's must be small and therefore, this term should be rather small therefore, by comparing with the tabulated value of a chi square distribution on k minus 1 degrees of freedom we can test whether H naught can be rejected or cannot be rejected if it is not true then this difference will tend to be large. So, the value of W will be large.

The test H naught is rejected, if W is greater than or equal to chi square k minus 1 alpha, now there may be a situation where F naught may not be completely known; that means, it may include certain parameter, if this include certain parameter then from the data, we can estimate that parameter also and then in place of e i we can say, we are getting e i hats and we can substitute there in case F naught x contains unknown parameter say theta is equal to theta 1, theta 2, theta n, then we can estimate theta from the sample and accordingly find pi i hat is equal to probability x belonging to i hat; that means, the estimate of this and e i hat is equal to n times pi i hat.

Then for large samples e i hats will converge to e i in probability or with probability 1.

(Refer Slide Time: 15:12)



$$W^* = \sum_{i=1}^{k} \frac{(O_i - \hat{e}_i)^2}{\hat{e}_i}.$$

W^* has asymptotically χ^2_{k-m-1} distn.

So we can use test: $W^* \geq \chi^2_{\alpha, k-m-1}$ rejection region.

We should not have π_i's very close to 0 or 1.

Similarly expected frequencies should not be below 5.

Examples 1. In a sale of 300 units of an item the following preferences for colours are observed for customers:

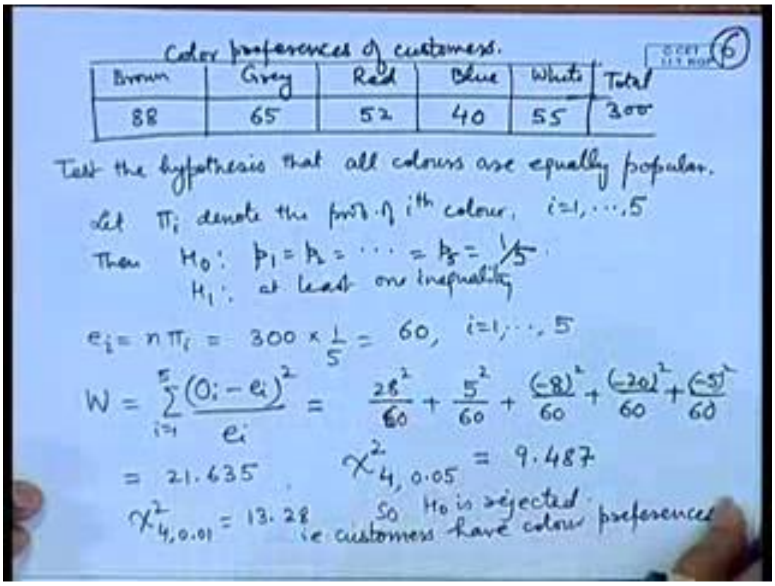
So, we can use say W star that is equal to sigma O i minus e i hat square by e i hat i is equal to 1 to k W star has asymptomatic chi square k minus m minus 1 distribution. So, once again we can use test as that W star greater than or equal to chi square alpha k minus m minus one this is the rejection region.

However there are certain precautions one should take while using the chi square approximation as the binomial approximation to the normal distribution is good when p is moderate; that means, it should not be close to 0 or close to 1 in a similar way, here the cell probabilities are say pi i’s. So, if either of the pi i’s is extremely small; that means, close to 0 or close to 1 then in that case, the expected frequency of that cell will become either too small or too large. If it is too large then for some other cell, it may become too small in that case this approximation is not good.

We have the following considerations, we should not or we can say we should not have pi i’s very close to 0 or 1, similarly expected frequencies should not be below 5. So, a practical consideration that has been done that there may be a case that. Firstly, we split the interval without knowing the probabilities, but when we actually calculated the probabilities and find that the expected frequencies are below 5 then what we do? We can merge some adjacent intervals. So, that the number of intervals becomes slightly less, but each cell frequency becomes more than 5.

This is a practical approach that is used here; let me explain this test through certain examples in a sale of say 300 units of an item the following preferences for colours are observed for customers. So, it may be like certain item of the type say for example, somebody is buying say car or somebody is buying a say 2 wheeler. So, we look at the colour of the car for example, so, out of 300 customers suppose we observe that brown; the colours available are brown, grey, red, blue and say white.

(Refer Slide Time: 19:21)



Color preferences of customers.

Brown	Grey	Red	Blue	White	Total
88	65	52	40	55	300

Test the hypothesis that all colours are equally popular.

Let π_i denote the prob. of i^{th} colour, $i=1,\dots,5$

Then H_0: $p_1 = p_2 = \cdots = p_5 = 1/5$

H_1: at least one inequality

$$e_i = n\pi_i = 300 \times \frac{1}{5} = 60, \quad i=1,\dots,5$$

$$W = \sum_{i=1}^{5} \frac{(O_i - e_i)^2}{e_i} = \frac{28^2}{60} + \frac{5^2}{60} + \frac{(-8)^2}{60} + \frac{(-20)^2}{60} + \frac{(-5)^2}{60}$$

$$= 21.635 \qquad \chi^2_{4,0.05} = 9.487$$

$\chi^2_{4,0.01} = 13.28$ So H_0 is rejected.

i.e customers have colour preferences

Out of 300 customers, we find 88 prefer brown, 65 prefer grey, 52 prefer red, 40 prefer blue and 55 prefer white, out of 300 colour preferences of customers. So, we want to test the hypothesis that all colours are equally popular; that means the customers have equal preferences for each of the 5 colours. So, if we want to frame a hypothesis in the form of test of goodness of fit, what we can do is the hypothesis is of the form that each cell has probability 0.2.

Let pi i denote the probability of i th colour for i is equal to 1 to 5, there are 5 colours here then we want to test that each of the is 1 by 5, now based on this assumption. So, here basically the cell or interval is actually the type here. So, brown is 1 type grey is another type, red is another type, blue is another type, white is another type. So, this is again a multinomial situation and here we are assuming the probabilities to be same in the null hypothesis.

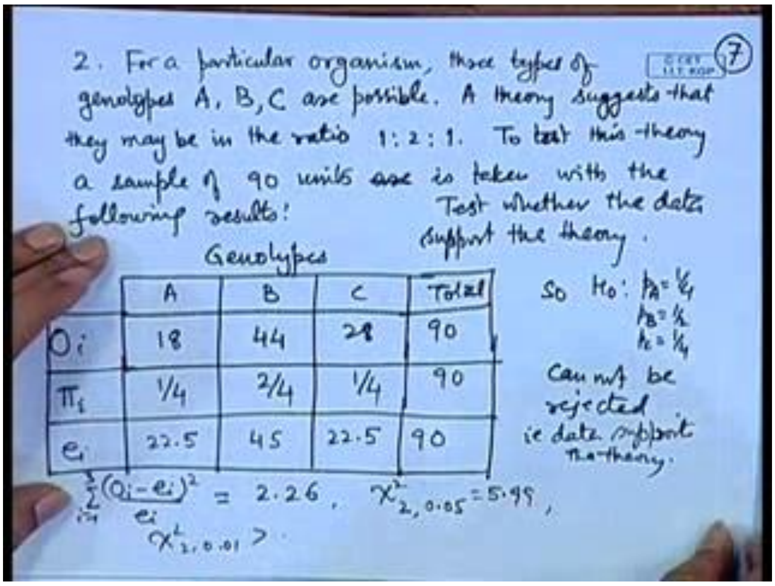
H 1 is at least 1 in equality. So, 1 the basis of this, we do the following calculations, we can calculate e i’s. So, e i is n pi i so here it is 300 into 1 by 5 that is equal to 60 each category has the same probability and therefore, each category will have the same expected frequency also. So, on the basis of this if we calculate W that is sigma O i minus e i square by e i i is equal to 1 to 5 then this is equal to 88 minus 60 that is 28 square by 60 plus 65 minus 60 second cell that is 5 square by 63rd cell is 52. So, it is minus 8 square by 60 plus 40 minus 60 is minus 20 square by 60 plus 55 minus 60 is minus 5 square by 60.

The sum can be easily evaluated, it turns out to be 21.635, now there are 5 categories, we also notice that the expected frequency of each cell is more than 5. So, the chi square assumption is valid therefore, we look at the value of chi square on 4 degrees of freedom suppose I consider the value at say 0.05 then from the tables of the chi square distribution 1 can find this value is 9.487, we may even look at say chi square 4.01 that is 13.28.

You can see that the calculated value of W that is 21.635 is bigger than this. So, H naught is rejected; that means, what is the conclusion the conclusion is that customers have preferences for the colours you can see the raw data here the observed frequency for brown is 88 which is almost more than twice the choice of blue colour, if we see the choice of grey that is much higher the choice of red blue and white is below. So, you can see that in specifically speaking brown and blue they cause major discrepancies here blue is say least favourable colour and brown is the most favourable colour here. In fact, if I have only 3 of this then they look almost nearby 65, 52 and 55 that is customers have colour preferences.

Basically what we have tested is something like a discrete uniform distribution and we conclude that the data does not follow a discrete uniform distribution.

(Refer Slide Time: 25:15)



7

2. For a particular organism, three types of genotypes A, B, C are possible. A theory suggests that they may be in the ratio 1 : 2 : 1. To test this theory a sample of 90 units ~~are~~ is taken with the following results: Test whether the data support the theory.

Genotypes

	A	B	C	Total
O_i	18	44	28	90
π_i	1/4	2/4	1/4	90
e_i	22.5	45	22.5	90

So H_0: $p_A = 1/4$
$p_B = 1/2$
$p_C = 1/4$

Can not be rejected
ie data support the theory.

$$\sum_{i=1}^{3} \frac{(O_i - e_i)^2}{e_i} = 2.26, \quad \chi^2_{2, 0.05} = 5.99,$$

$$\chi^2_{2, 0.01} >.$$

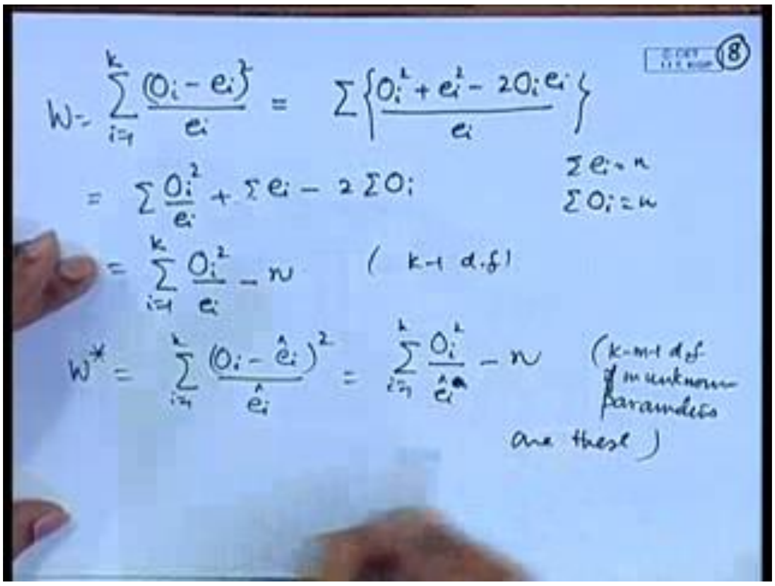
Let us take another example for a particular organism, 3 types of genotypes A B and C are possible, a theory suggest that they may be in the ratio say 1 is to 2 is to 1. Now to test this hypothesis to test this theory a sample of 90 units is taken with the following results.

Genotypes A B C; it is observed that out of 90 units 18 had genotype A, 44 had genotype B and 28 had genotype C, the total is 90. So, now, we want to test whether the data supports the theory. So, for this once again we have 3 categories the probabilities of each category let me call it pi i. So, 1 is to 2 is to 1. So, this probability is 1 by 4, this probability is 2 by 4 that is half and this probability is 1 by 4.

Expected frequency this is observed frequency we can actually do the calculations in the form of a table here. So, here you see if the probability of the genotype A is 1 by 4, the total number of units is 90. So, the expected frequency for that will be 90 by 4 that is 22.5, here it will become 45, here it will become 22.5 that is total is 90.

Based on this, one can carry out the calculations sigma O i minus e i square by e i i is equal to 1 to 3. So, for example, here it will become 4.5 square divided by 22.5 plus 144 minus 45. So, 1 square by 45 and 28 by minus 22.5 that is 5.5 square by 22.5 so one can look at these calculations this turns out to be 2.26. So, one can easily compare with the chi square value on 2 degrees of freedom suppose we look at 0.05 then this is 5.99 and of course, if I look at say chi square 2 at 0.01, this is going to be larger than this. So, we cannot reject H naught; that means, H naught that P A is equal to 1 by 4 P B is equal to half P C is equal to 1 by 4 cannot be rejected; that means, the data supports the theory that the genotype or in the proportion 1 is to 2 is to 1, here 1 point about this calculation also this formula which we have given here sigma O i minus e i square by e i.

(Refer Slide Time: 30:06)



8

$$W = \sum_{i=1}^{k} \frac{(O_i - e_i)^2}{e_i} = \sum \left\{ \frac{O_i^2 + e_i^2 - 2O_i e_i}{e_i} \right\}$$

$$= \sum \frac{O_i^2}{e_i} + \sum e_i - 2\sum O_i \qquad \sum e_i = n \quad \sum O_i = n$$

$$= \sum_{i=1}^{k} \frac{O_i^2}{e_i} - n \qquad (k-1 \text{ d.f})$$

$$W^* = \sum_{i=1}^{k} \frac{(O_i - \hat{e}_i)^2}{\hat{e}_i} = \sum_{i=1}^{k} \frac{O_i^2}{\hat{e}_i} - n \qquad (k-m-1 \text{ d.f. if } m \text{ unknown parameters are there})$$

One can actually have an alternative form for this; we can consider expanding this term. So, it is O i square plus e i square minus twice O i e i divided by e i that is equal to sigma O i square by e i. Now the next term here is e i and then summation. So, summation e i is actually n minus twice summation O i. So, summation e i is n and summation O i is also n. So, this becomes simply sigma of O i square by e i minus n i is equal to 1 to k.

This is an alternative formula for W star for W similarly, if I am considering W star that is sigma O i minus e i hat a square by e i hat then once again this can also be written as sigma O i square by e i hat square minus sorry, e i hat minus n, here degrees of freedom are k minus n and here the degrees of freedom are k minus m minus 1, if m unknown parameters are there many times this expression is easier to calculate.