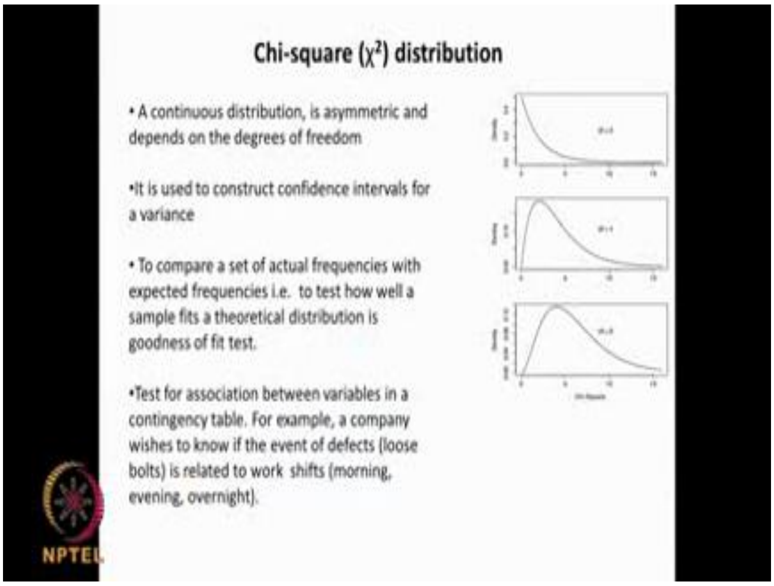
Biostatistics and Design of Experiments
Prof. Mukesh Doble
Department of Biotechnology
Indian Institute of Technology Madras

Lecture - 22
X2 distribution/test

Welcome to the course on Biostatistics and Design of Experiments. We will talk about chi square distribution then χ^2 CHISQUARE test these very, very important especially when you are looking at data where we are talking about some sort of binomial data; yes-no, live-dead, one-zero and that sort of situation.

(Refer Slide Time: 00:37)



Now, what is Chi square distribution? These also a continuous distribution, just like your normal, but then it is somewhat related to binomial at some point, we will come to that actually. It is asymmetric and depends on the degrees of freedom. For example, if we have very low degree of freedom it looks like this. As the degree of freedom increases, you end up slowly, slowly having skewed and so on actually. So, it is used to construct confidence interval for variance. Generally, Chi square is used for variance ratios. It can be used to compare a set of actual frequencies with expected.

So, I expect certain number of child mortality in a village, but the current data is this much, is it significant or not significant. I expect to have a certain number of defects when I manufacture bolts, but i see so many defects in a day. But so, is it significantly different or not? Such sort of situation actually. So, it is actual frequencies versus the expected frequencies. So, it is like we are trying to fit with a theoretical distribution and then check for the goodness of the fit actually. So, I expect a particular sort of values, but I find this. Like I expect 60 to 70 % marks of a 10 students in a class and 70 to 80 of another 10 students in a class and above 90 % another 10 students. But I find different numbers in that range of 60 to 70, 70 to 80, 90; now is that really a significant difference?

It also looks at association between variables in a table. We will talk about this contingency table later on actually. So, I make some equipments, some parts using three or four different equipments. I find defects in each of the equipments. So, defect - non defect, defect - non defect, equipment 1, equipment 2, equipment 3. Is there a relationship between the defects and equipment? So, I can use a Chi square test for that. This is very very useful especially in clinical trials, you may be running lot of samples on HPLCs and some HPLCs may give very accurate result. So, i it is acceptable. Some HPLCs sometimes they give accurate sometimes they give non accurate which is not acceptable.

So, I may have 100 samples injected in HPLC 1; I may have say 30 rejected and 70 accepted; I may have 100, 110 samples injected in HPLC 2; I may have 40 rejected and 70 accepted and so on actually. So, is there are association between the rejects and the instrument? For that sort of situation, I can use something called the Chi square distribution.

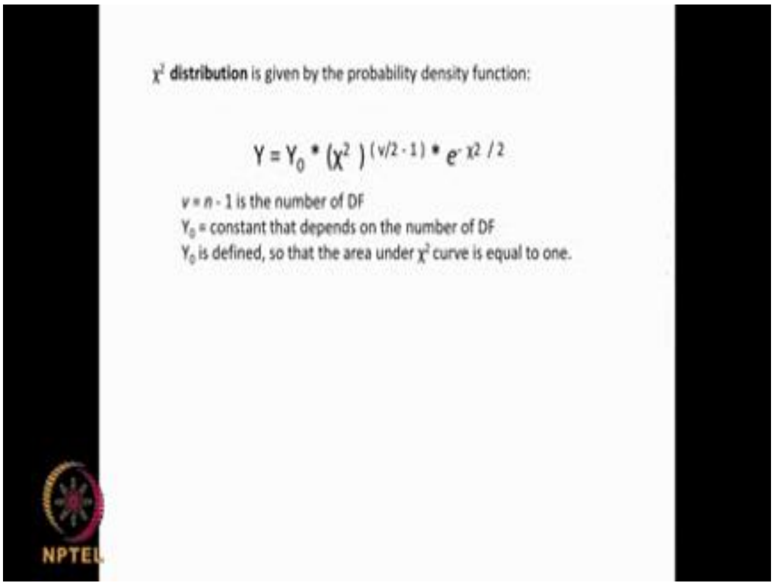
So, for example, a company wishes to known if the events of defects is related to work shift. This is very common in industries you know like in industries people who are come in the morning shift like 8 AM to say 4 PM and then there will be second shift which could be 4 pm to night midnight 12 o’clock and then there could be a third shift which could be night midnight 12 o’clock right up to 8 AM. Now each of the, each of the workers in the each of these shifts are manufacturing a particular part like bolts for example. Now, you find some failures in some shifts, some successes in some shifts, morning shift, after evening shift, night shift.

Now is there an association you want to know? Because, it is always considered that people working in the night shifts may always make mistakes. So, mostly defects come in the night shifts. So, is there an association between the defects on the failures as against to the shift that is time when which workers work. So, this type of situation also we use this Chi square distribution and Chi square test.

So, in each of these it is very very powerful and for you to look at Chi square type of distribution. And of course, as i said in biological situation, it is very useful because you are telling we expect a mortality of a say 10 % and then we find in few villages different numbers. Is it significantly different? In that sort of situation, we can use this type of Chi square test actually.

You are testing a drug, it shows successes it shows failures. You are testing a placebo it show successes it shows failures. Now is there a difference between the placebo and the drug or there is no difference? We use the Chi square test actually. So, it is a continuous distribution and it gets skewed like this at very low DF it is, it looks like this and as the DF keeps increasing it slowly becoming more uniform and at very large DF you may end up having some sort of a nice smooth uniformly distributed picture.

(Refer Slide Time: 06:25)

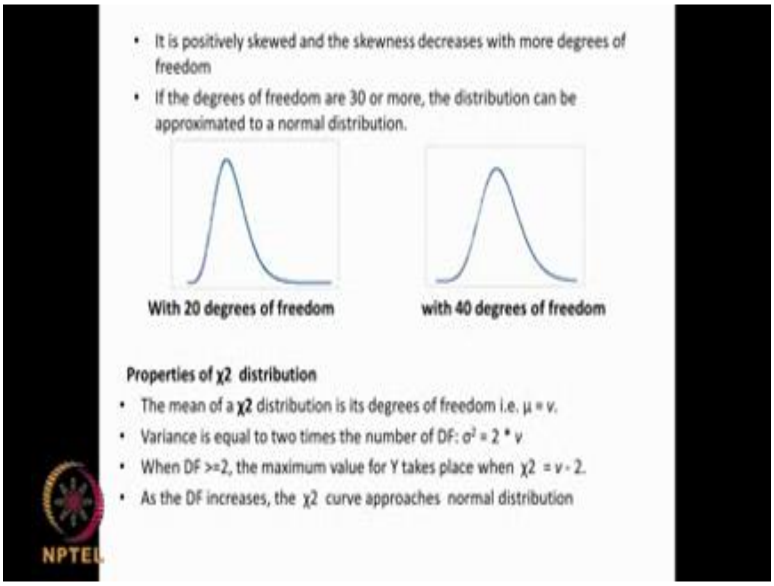


So the equation for this looks like this. Y,

$$Y_0 * (\chi^2\ )$$

value, this is the probability density function. This is nothing but the degrees of freedom. So, this is how these equation looks like. So, $Y = Y_0$; Y_0 is a constant. Constant that depends on the number degrees of freedom, Chi square is the probability density function and that is how you calculate, ok?.

(Refer Slide Time: 07:06)



So, it is generally it is positively skewed like i showed you you know positively skewed generally. But as the degrees of freedom keeps increasing the distribution can be approximated to normal. Look at this, this is the 40 degrees of freedom, it looks almost like a normal distribution. So, what are the 20 degrees you can still see these skewness here? That is why it is generally believed that if you have 30 or more degrees of freedom it will approximate to a normal distribution. So, the mean of a Chi square distribution is the degree of freedom. That is

$$\mu = \nu$$

mu is equal to this value V or Nu it is called actually, that is a DF. Variance is equal to 2 times the number of degrees of freedom. So,

$$\sigma^2 = 2 * v$$

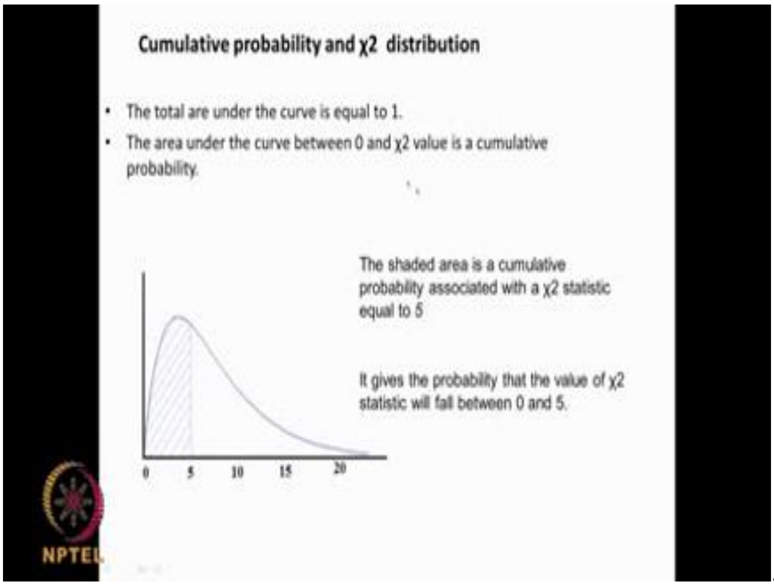
. So, in a Chi square distribution so the mean is equal to the number of degrees of freedom and the variance will be 2 times that. When DF is > or = 2, the maximum value for Y takes place when

$$\chi 2 = v - 2$$

do you understand? As the degree of freedom increases; that means, the maximum value, occurs for the Y when Chi square = DF - 2.

Imagine, I have a degrees of freedom as three. So, maximum value will occur at 3 -2, 1. As the degree of freedom increases, Chi square curve approaches normal distribution as you can see from here. So, the Chi square will have a mean of degrees of freedom, the variance will be 2 times the degrees of freedom.

(Refer Slide Time: 09:00)

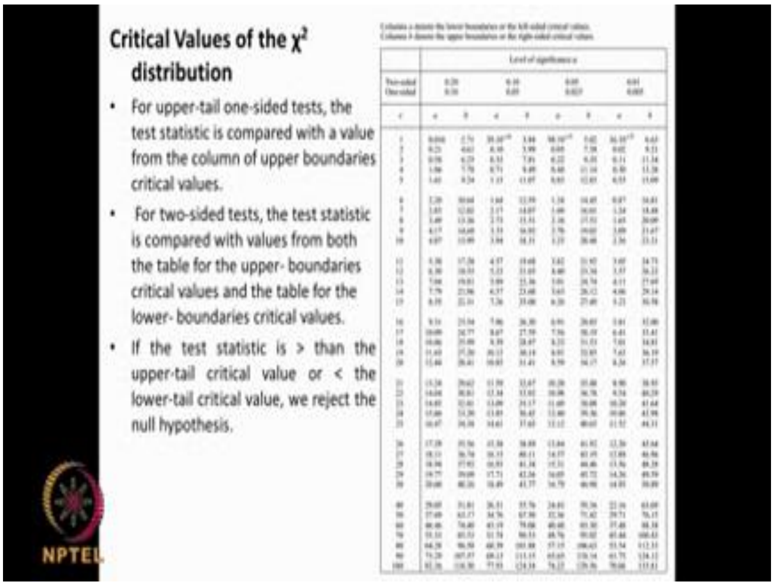


.

We can get the cumulative probability also the total area under the curve is = 1. Generally it is put into there, and the area under the curve between 0 and Chi square value is this cumulative this is the area. So, the distribution generally looks like this with a long tail because as i said, it is got a positive skewness. So, this is how it is look like. Here it rises very fast and then it falls down and then it keeps on going.

So, if you are talking about cumulative for $\chi^2 = 5$, then this is the entire area; this is entire area. But the total area under this curve = 1. So, it gives the probability that the value of Chi square statistics will fall between 0 and 5. So, just like your t distribution, normal distribution; so, if $\chi^2 = 5$, the probability the area under the curve is the probability. Because this whole area is 1; whole probability is 1. So, this gives you the probability that your χ^2 will lie between 0 and 5, do you understand? χ^2 lying between 0 and 5 is given by the shaded area. If we take the whole area, the probability is 1 and the area is also 1.

(Refer Slide Time: 10:21)



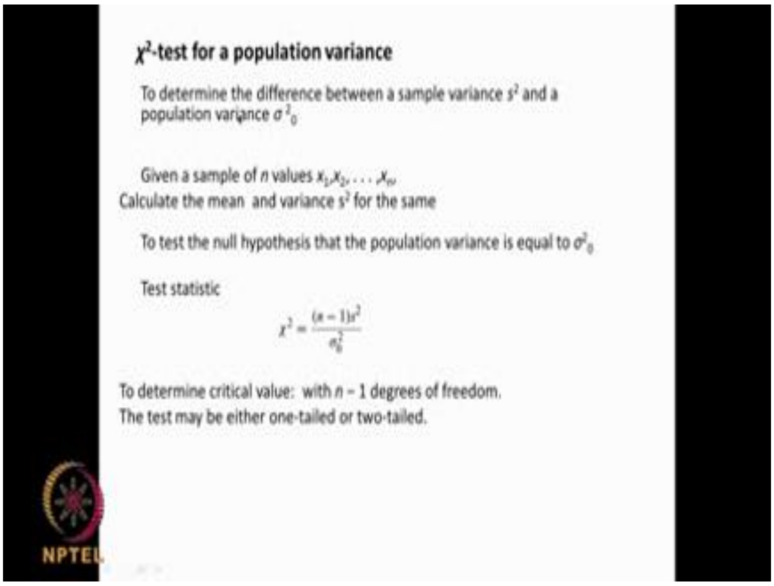
So just like t table, z table, F table, we also have a table for χ^2 and you have a here the degrees of freedom and then here we have the α and of course you use as you can see two sided and one sided; so one sided will be half of this 0.05 means 0.025; 0.012 sided means 0.005; obvious right? So, this is how it looks like.

So, for different degrees of freedom, will have 2 columns here. This A is the lower boundary that is on the left hand side critical value. B is the upper boundary right hand side. On the right hand side and that even is upper boundary. So, we see there is lot of difference in the magnitude. Why is that? That is because the way the shape of the curve looks like. So, if i put in here a line here for example, and a line here for the area, because it is raising very sharply, we will have very small value because it is falling down very slowly you will have a very large value. That is why if you look here, the A and B, you can see small number and on the B is a very large number, small number, very large number. So, like that it goes actually. Do you understand?

And this top one relates to two sided, bottom one relates to one sided. So, we use it for two tail test and one sided test as obvious, right?. So, we prepare the test statistics and then we compare from both the table for the upper boundaries and the table for the lower boundaries as well. The test statistics which we use, I will tell you what the test statistic is. Is greater than the upper tail value or less than the lower tail critical value, we reject the null hypothesis. So, if I am looking at this and if my test statistics for χ^2 comes out to be say for example, larger than 2.71, then I reject the null hypothesis. For 1 degree of freedom, for probability of two sided 0.2, I am saying 2.71.

So, for 1 degree of freedom, probability two sided 0.05, the if I get more than 5.02 if my χ^2 calculated is more than 5.02, then I reject the null hypothesis; you understand? Same thing here, 3.84 for a one sided test. If I get more than 3.84 when I calculate the χ^2, I reject the null hypothesis. This is how you do and these this as you go down we are having more degrees of freedom coming into the picture. So, this χ^2 test can be used for contingency table, it can be used for comparing two different distributions, it can be used for a comparing expected versus observed and so on. We will look at each of them slightly more in detail

(Refer Slide Time: 13:32)



So, you can use $\chi 2$ for a population variance. So, I have a sample variance, I want to know whether it comes from that population. So, I have a sample n values, x1, x2, x3 up to xn. So, I can calculate mean and variance, s^2 for this one. Now I want to know whether this s^2 comes from the population of σ_0. So obviously, what is the test statistics here?

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}$$

. n is the number of data points. So, this is the test statistic. And then we from based on the number of degrees of freedom based on your level of test, you take the appropriate table Chi square value, and see whether this value is greater than the table value? In that case we can reject the null hypothesis saying that null hypothesis they come from the same population, alternatives they do not come from the same population. And we can use either one tail or two tail test actually. So, the null hypothesis will always be status quo they come from the same population; that means, the sample comes from the same population ,alternate will be sample does not come from the same population, ok?.

(Refer Slide Time: 14:50)

So, let us look at this simple problem, just understanding. So, the null hypothesis they come.

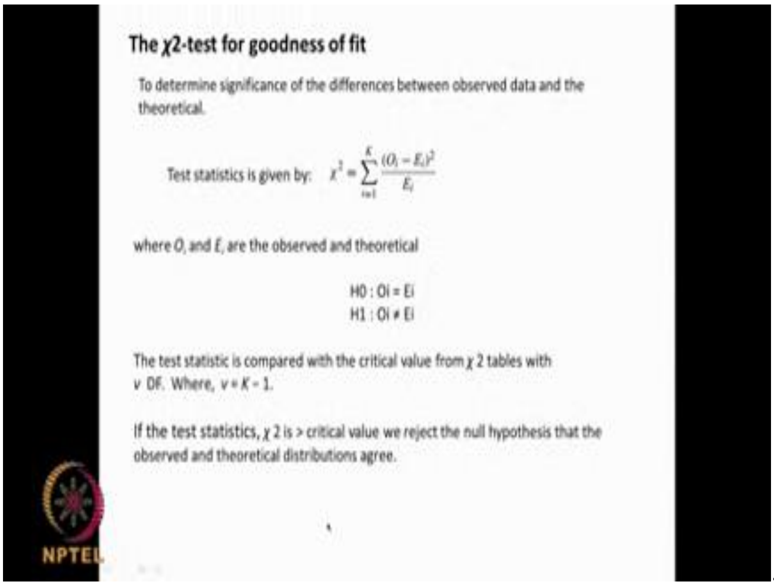
$$\sigma^2 = \sigma_0^2$$

alternate is they are not. What is the test statistics? Like I said n - 1. So, I have a sample right? So, $\frac{(n-1)s^2}{}$. is your standard deviation divided by σ, not square. So, sigma, sigma is given by 4, number of data point is 17, s^2 is 7. So 16, so this will be 16 * 4, 16 * 4 ÷, sorry 16 * 7 ÷ 4. 16 because the number of data point is 17, 16 * 7 ÷ 4. 4 is your σ. So, 6 16 / 4 is 4. 4 *7 is 28.

So, from the test statistics we get $\chi^2 = 28.0$. So, we have a 17 data points. So, 16 degrees of freedom, 16 degrees of freedom is in this row and we all suppose we look at two sided 95 %; that is this place, two sided 95 % is this place. One sided 95 % will be this place, this this column. So, two sided 95. So, we go here; it is 28.85. So we calculated as 28.0, see we accept the null hypothesis. For a one sided test, for a one sided test, for two sided test.

Suppose we take, yeah, 28 point. Suppose we take the other one: that is we take one sided test instead of a two sided 95 %; then it will become 26.3 here. So, we can reject the null hypothesis. So, if it is a two sided, 95 % it will become 28.85 the test statistics is 28; so obviously we cannot reject the null hypothesis. If it is a one sided 95 %, then you are table is table is 26.3; your test statistics is 28. So, we reject the null hypothesis, depending upon what what is your hypothesis that's being developed.

(Refer Slide Time: 17:48)



.

Now, we can use this for testing the goodness of fit also. So, what do we do? We look at the observed, we will look at the expected. Expected is your theoretical value. So,

$$\chi^2 = \sum_{i=1}^{K} \frac{(O_i - E_i)^2}{E_i}$$

and sum it up. So, i have 10 data points means, I do 10 summation. Observed -expected / expected here you have a square term actually. So, you are null hypothesis

H0 : Oi = Ei

, alternate will be observed

$H1 : Oi \neq Ei$

So, remember your test statistics is compared with the then, the the null hypothesis is

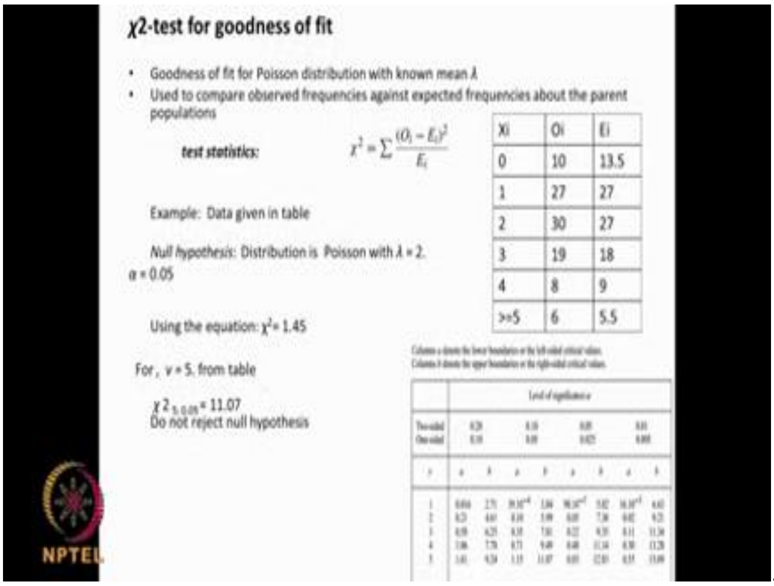
$H0 : Oi = Ei$,

alternate is

$H1 : Oi \neq Ei$

. And then after that what you do? You go to your table if your test statistics > the critical value we reject the null hypothesis that the observed is equal to theoretical distribution otherwise we accept the null hypothesis.

(Refer Slide Time: 19:00)



.

Let us again look at a simple problem. We are comparing observed frequencies against expected. So, there is an observed frequency. So, with 0 value there are 10, 1 value 27, 2. So, we can say the there are 10 students with certain group of marks, 27 students with a certain a group, 30 students with some other group of marks, 19 students with some other group, 8 students with some other group of marks, 6 students with, but we expect this.

So, we want you to look at the hypothesis whether they come from the same or not. So, what we do?

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

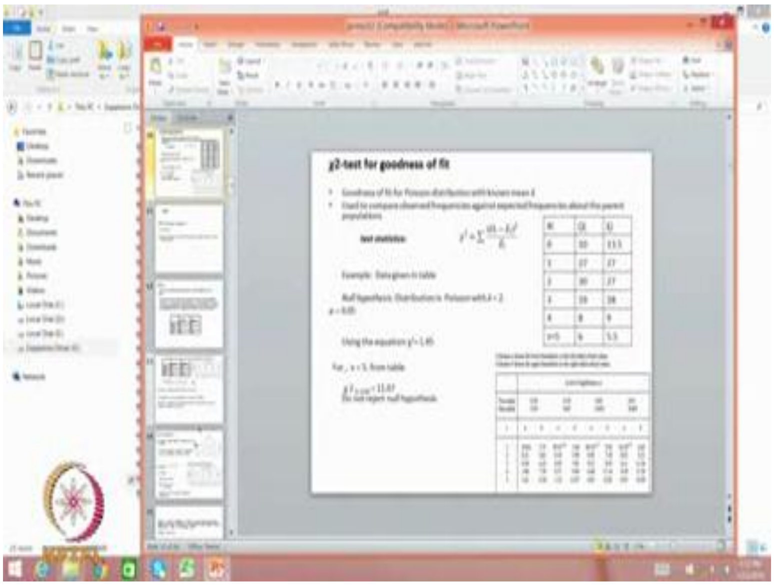
that is, this minus this square ÷ expected like that you keep on adding each one of these. When we do that, we get 1.45. Now, if you go to your table for 3, 4, 5, 6; so 5 degrees of freedom we go to your table, we go to our table for 5 degrees of freedom. So, if it is one sided, we take for 95 % it is 11.07. So, we and we calculated as 1.45. So, we do not reject the null hypothesis. Do you understand this problem?

So, here what we say is, observed = expected, that is null hypothesis. Observed ≠expected, that is the alternative hypothesis. So obviously, the Chi square test statistics which we calculate should be greater than this value. Here we are taking for 5 degrees of freedom and for one sided 95 %.

(Refer Slide Time: 20:56)

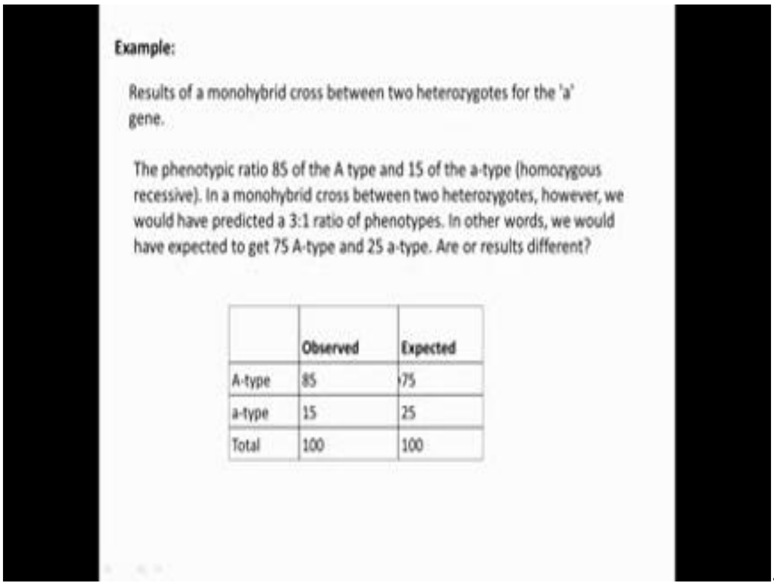
Excel also has a function it is called CHITEST. That is actual range comma expected range it gives you the statistics actually. Let us look at this CHITEST thing also in excel.

(Refer Slide Time: 21:08)



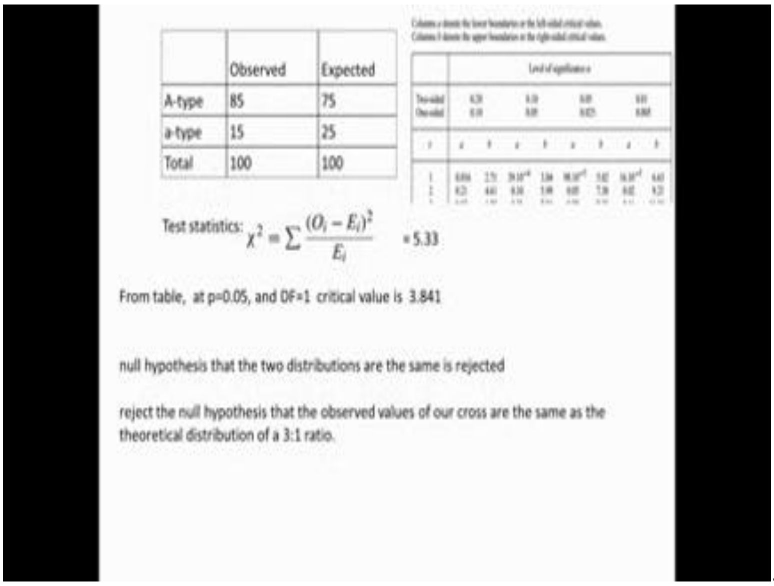
So, we go to excel, so we say 10 we will put in the data 27, we will put in the data 30, then we put is as 19, 8, 6; 13.5, 27, 27, 18, 9, 5.5. So, we do Chi; see you have this CHITEST comma comma. So, it is giving amount, actually this is giving you the probability 0.918. So obviously, the value is very very large, so there is no chance of rejecting the null hypothesis and that is what we get from our thing also here. So, it gives you the from the statistics; it gives the statistics here so it gives you 0.91. So, there is no chance of rejecting the null hypothesis from this, there are set. Let us look at some more problems; again looking this is these are called contingency tables, where we have a expected and observed coming into picture. So, we look at some problems in this particular area.

(Refer Slide Time: 23:08)



So, results of a monohybrid cross between 2 heterozygotes for the A gene, so we are looking at it. The phenotypic ratio 85 of the A type and 15 of the small a type, this is the ratios. In a monohybrid cross between 2 heterozygotes; however, we would have predicted a 3 is to 1; 3 is to 1 is nothing, but 75, 25. Now, we want to know whether these results are very different. So, our null hypothesis will be; there is no difference between these two observed and expected. So, if you remember this; they are same and the alternate is they are not the same so what we do?

(Refer Slide Time: 24:04)

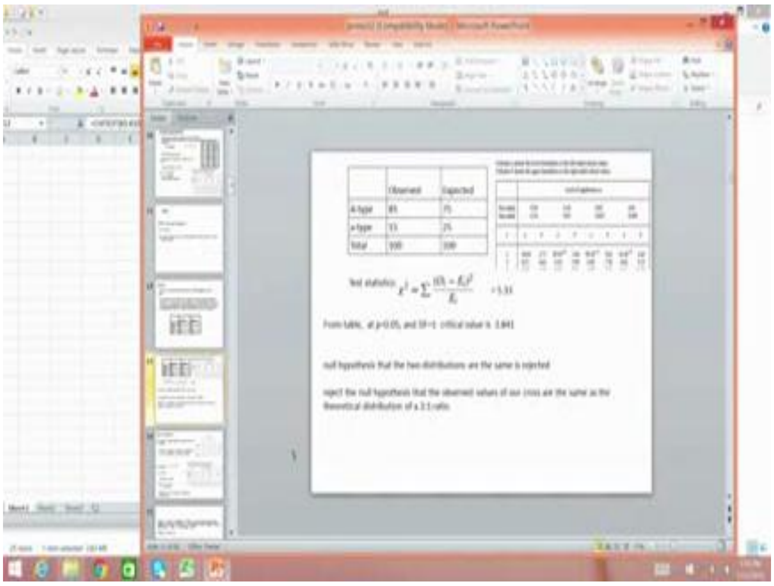


We use the same formula

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

; add these two. So, we end of with 5.33. From the table, because it is 2 / 2; it will be 1 degree of freedom, so we go to 1 degree of freedom; it is 3.841, so; obviously, the null hypothesis same is rejected. So, we reject the null hypothesis observed = expected. So, once you reject the null hypothesis, that the observed values over cross are the same as the theoretical distribution of 3 : 1. So, generally we expect three : one, but we find phenotypic ratio of 85, 15, which is not the same, we understand? So, we can again use the excel also.

(Refer Slide Time: 25:04)



So, we can use the excel also 85, 15 and 75, 25, 75, 25. So, CHITEST; actual divided by expected, we get 0.02. So obviously, p value is 0.02, so obviously, it is statistically significant. So, you see this from CHI TEST, we can also calculate CHITEST command which is available in excel. We can that is this one, we can calculate these particular value. Now, there is a command called CHI INVERSE, what is CHI INVERSE? We give the probability; we give the probability say for example, I give probability and 1 degree of freedom. So, I need to get 3.84, as you can see here so it gives you this Chi inverse command is almost like your table. So, Chi CHI INVERSE command, 3.84. You understand? At 1, at 0.05 with 1 degree of freedom, as you can see here; 0.05 one sided comma 1; 1 degree of freedom, gives you 3.841.

So, CHI INVERSE is almost like your, it gives you the table; where as this one gives you the the CHI TEST gives you the probability. So, in the CHI TEST, we give the observed and then we give the expected on one side; we give the observed on the other side. That way, sorry, expected on one side, observed on one another side. So, the actual and the expected that will give you the probability; whereas the CHI INVERSE, given the probability and the degrees of freedom, it gives you the Chi square value, as shown in the table like point I mean 3.84 for example, 3.84. So, that is given in the CHI INVERSE.

So both are very useful type; of a things to have when you are talking about. So, Chi distribution, this is an older version for excel 2007. So, we have CHITEST, which if I give the observed and expected will give me the probability; where as in the CHI INVERSE, when I give the probability and degrees of freedom it will give you the Chi square value. So, both these excel commands are very useful also, but you can also do it by simple calculator. Because, it is not very difficult for one to do, the formula is very simple

$$\frac{(O_i - E_i)^2}{E_i}$$

, observed minus expected, square it up; ÷ expected and then plus Oi – Ei^2; ÷ Ei. So, so simple to do also, it is not very difficult for one to perform this type of calculations. So, we will continue more of this, Chi square distribution again in the next class also.

Thank you very much.

Key words: Confidence interval, variance ratio, frequency, variables, degree of freedom, probability, cumulative probability, mean, standard deviation, contingency table, chi square, chi test, chi distribution