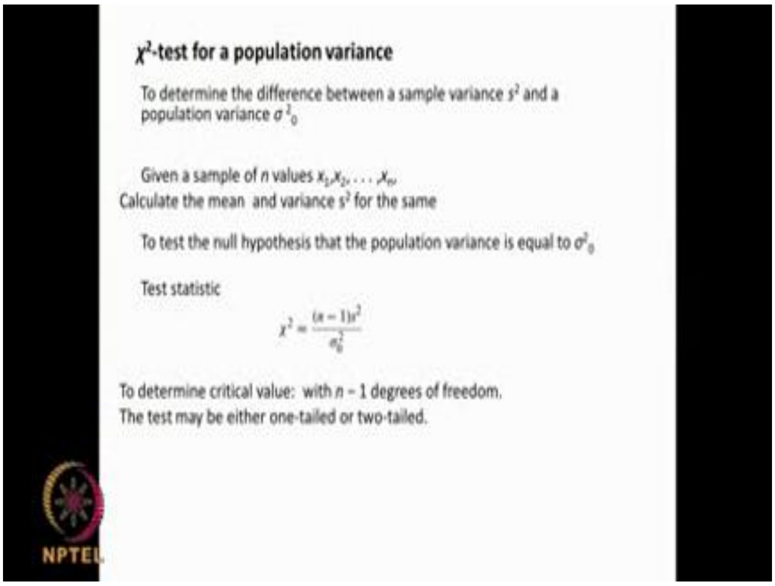
Biostatistics and Design of Experiments
Prof. Mukesh Doble
Department of Biotechnology
Indian Institute of Technology, Madras

Lecture - 24
χ^2 test

Hello everyone, welcome to the course on Biostatistics and Design of Experiments. We will continue on the topic of χ^2 test. It is very important test, it is almost like binomial type of test- fail-pass, yes-no, naught and so on actually.

(Refer Slide Time: 00:33)



χ^2-test for a population variance

To determine the difference between a sample variance s^2 and a population variance σ^2_0

Given a sample of n values $x_1, x_2, \ldots, x_n$,
Calculate the mean and variance s^2 for the same

To test the null hypothesis that the population variance is equal to σ^2_0

Test statistic

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

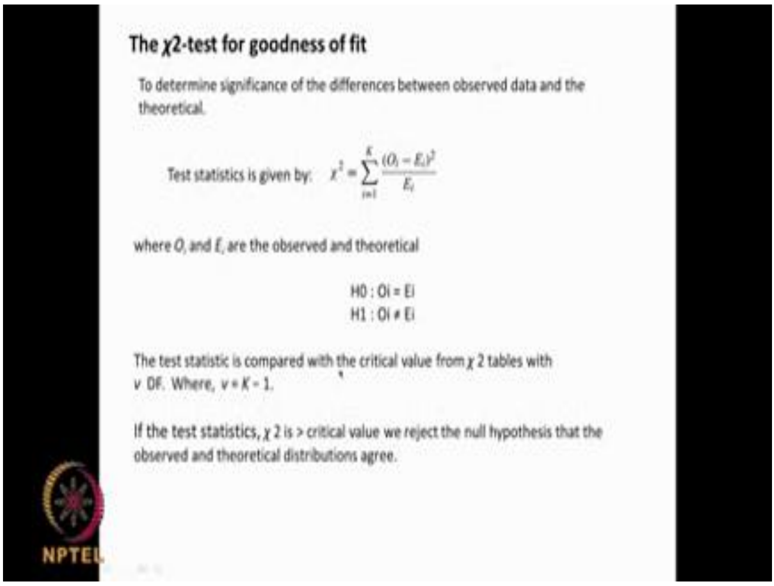
To determine critical value: with $n - 1$ degrees of freedom.
The test may be either one-tailed or two-tailed.

We can use this for looking at sample versus population. Suppose, we have a sample variance of $s^2$and a population variance of • σ^2, we can use this to testing the null hypothesis that the population variance = • σ^2. That means, we want to say whether the sample comes from the same population. When we have a set of samples x1 to xn we are calculating a mean and a variance, that will called it s^2 and then the original the population variance is • σ^2. Then we calculate something call the test statistics for the

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

. Here, the degrees of freedom will be n - 1, so s^2 will be the variance for the sample, • σ^{2}_{0} is the variance of the population and we are trying to see whether the sample comes from the same population. There is another table just like other tables, so we can calculate the table • χ^2 and then either we reject or accept null hypothesis based on comparing the test statistics with the table value.

(Refer Slide Time: 01:47)



The χ2-test for goodness of fit

To determine significance of the differences between observed data and the theoretical.

Test statistics is given by: $\chi^2 = \sum_{i=1}^{K} \frac{(O_i - E_i)^2}{E_i}$

where O_i and E_i are the observed and theoretical

H0 : Oi = Ei
H1 : Oi ≠ Ei

The test statistic is compared with the critical value from χ 2 tables with v DF. Where, $v = K - 1$.

If the test statistics, χ 2 is > critical value we reject the null hypothesis that the observed and theoretical distributions agree.

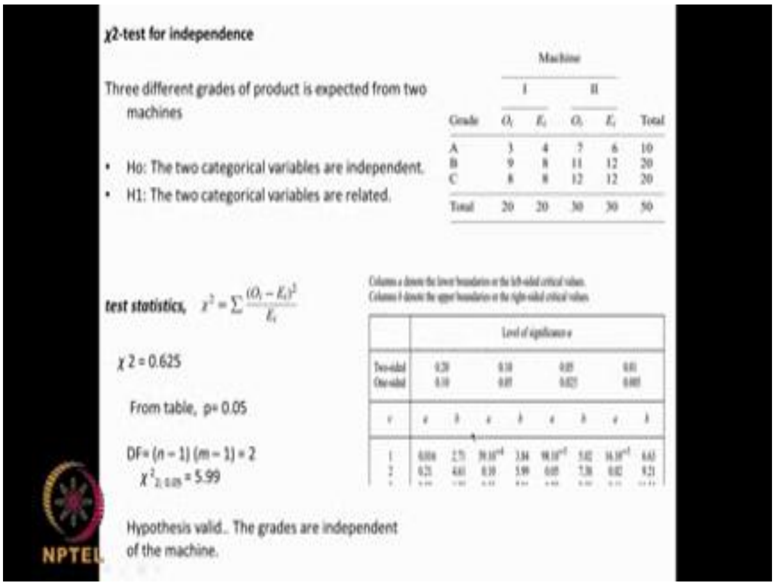
.

We can use it for testing goodness of fit. That means, you observe something where as you expect something, so obviously, is the observation is different from your expected or the observation is equal to the expected. The test statistics will be like this,

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

some of summation over K values. So, the null hypothesis here will be observed = expected and the alternate will be observed is not equal to expected. Again, we look at the degrees of freedom which will be K - 1 that is total number of data points and if the test statistics is > the table or critical value we reject the null hypothesis.

(Refer Slide Time: 02:35)



χ2-test for independence

Three different grades of product is expected from two machines

- Ho: The two categorical variables are independent.
- H1: The two categorical variables are related.

	Machine				
	I		II		
Grade	O_i	E_i	O_i	E_i	Total
A	3	4	7	6	10
B	9	8	11	12	20
C	8	8	12	12	20
Total	20	20	30	30	50

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

χ 2 = 0.625

From table, p= 0.05

DF= (n − 1) (m − 1) = 2

$\chi^2_{2,\,0.05}$ = 5.99

Hypothesis valid.. The grades are independent of the machine.

Column a denotes the lower boundaries or the left-sided critical values.
Column b denotes the upper boundaries or the right-sided critical values.

	Level of significance α							
Two-sided One-sided	0.20 0.10		0.10 0.05		0.05 0.025		0.01 0.005	
ν	a	b	a	b	a	b	a	b
1	0.016	2.71	$3.9 \cdot 10^{-3}$	3.84	$9.8 \cdot 10^{-4}$	5.02	$1.6 \cdot 10^{-4}$	6.63
2	0.21	4.61	0.10	5.99	0.05	7.38	0.02	9.21

And then we can also use it for looking at independence that means, we are performing some experiments and in one machine then we are performing some experiments in another machine. So, the results we get are they some of co-related to the machine or they were independent of machine. For example, this happens quite a lot when we are doing some analytical work on say HPLC, I have 2 HPLC, HPLC 1 and HPLC 2. I get some results in HPLC 1 I get some results in HPLC 2.

Now I want to see whether the results are independent of the machine or the results are dependent of the machine. This is a very important type of knowledge especially in clinical trials. So, that is χ 2 test for independence. Again, here we the test statistics will be

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

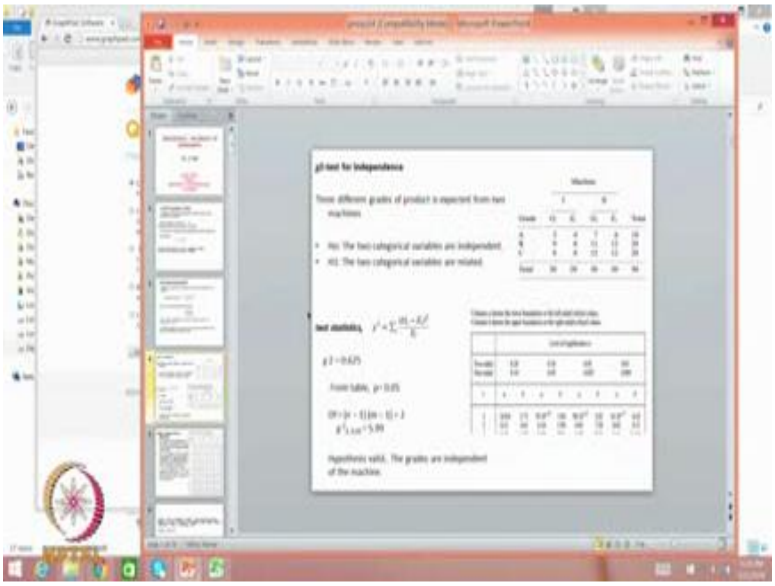
. So, what will be the null hypothesis here, the 2 categorical variables are independent. The 2 variables are independent of each other or the alternate hypothesis will be the 2 categorical variables are related to each other. So, the test statistics will be Oi-Ei For example, let us take this problem we have 2 machines, each machine makes 3 grades of product A, B, C are the grades, machine 1 machine 2.

These are the observed values 3, 9, 8 that means, machine 1 is making a 3 of grade A, 9 of grade B and 8 of grade C. Whereas we expect 4, 8, 8 from machine 1. Similarly, on machine 2 we see observed is 7, 11, 12 whereas we expect 6, 12, 12. The null hypothesis is the grades that it produces are independent on the machine on which they were tested or the alternate will be the two variables that is the grades and the machines or correlated with each other. So, we perform this test statistics. So,

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

. Let me show you how to do it in excel.

(Refer Slide Time: 04:53)



χ2 test for independence

Three different grades of product is expected from two machines

- H0: The two categorical variables are independent
- H1: The two categorical variables are related

	Machine				
	I		II		
Grade	O_1	E_1	O_2	E_2	Total
A	3	4	7	6	10
B	9	8	11	12	20
C	8	8	12	12	20
Total	20	20	30	30	50

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

$\chi 2 = 0.625$

From table, $p = 0.05$

$DF = (r-1)(m-1) = 2$

$\chi^2_{2,0.05} = 5.99$

Hypothesis valid. The grades are independent of the machine.

Let us do it on excel, it is quite simple. What you do? You put in all the data and then this is observed say 3, 9, 8 correspondingly we have 4, 8, 8, then we have a 7, 11, 12 then correspondingly we have 6, 12, 12. This is observed, this is expected. What do we have to do,

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

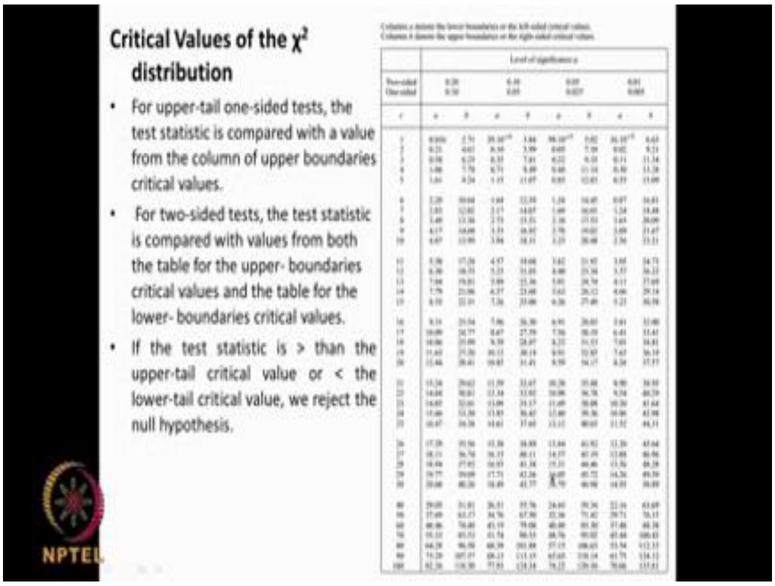
observed minus expected, square it divided by expected, so this is equal to this minus this square divided by expected is the second column, do you understand?. Oi - Ei 3 - 4 is 1^2, ÷ expected 1 / 4 that is 0.25. We do the same thing for all the data and then we add them up ,summation comes out to be 0.625. So, summation of this

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

summation comes to be 0.625.

Now let us go to the table, there is a table for critical value for • $\chi 2$ just like a table for your t ,z ,F and so on.

(Refer Slide Time: 06:32)



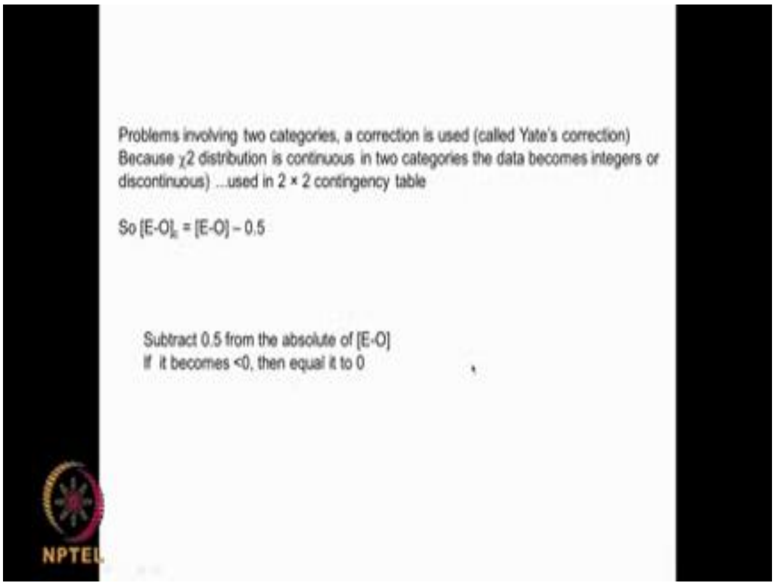
Critical Values of the χ^2 distribution

- For upper-tail one-sided tests, the test statistic is compared with a value from the column of upper boundaries critical values.
- For two-sided tests, the test statistic is compared with values from both the table for the upper- boundaries critical values and the table for the lower- boundaries critical values.
- If the test statistic is > than the upper-tail critical value or < the lower-tail critical value, we reject the null hypothesis.

So, let us look at, say it has got two-sided alpha and one-sided alpha. If you are looking at one-sided alpha here, so how many experiments are there, totally we have sorry, how many we have say 3 grades, so the degrees of freedom is 2, we have 2 machines so degrees of freedom is 1, so 2 * 1 is 2. So, the degrees of freedom = 2 in this case. The degrees of freedom is 2 and then we are going to looking into one-sided 95 %, So you get it as 5.99 here. The table value is 5.99, so the test statistics is 0.625. So, null hypothesis is valid, what is it?, the grades are independent of the machine from which it is manufactured. It is a very important study.As I said in clinical trials, samples are analyzed in different HPLC's and from different locations. So, I want to know whether the results are independent or you get one set of results if you try it out on one particular HPLC. That way this is a very useful type of test, one is to perform actually.

So, this is the table I talked about so we for a two-sided you have it here, for one-sided you have it here. So one-sided 95 % is here, two-sided 95 % is here, a denotes the lower boundaries and b denotes the upper boundaries and this is the degrees of freedom. We can use this table for different degrees of freedom and depending upon one-sided 95 % or a two-sided 95 % we use the corresponding columns here.

(Refer Slide Time: 08:25)



Problems involving two categories, a correction is used (called Yate's correction)
Because $\chi2$ distribution is continuous in two categories the data becomes integers or discontinuous) ...used in 2 × 2 contingency table

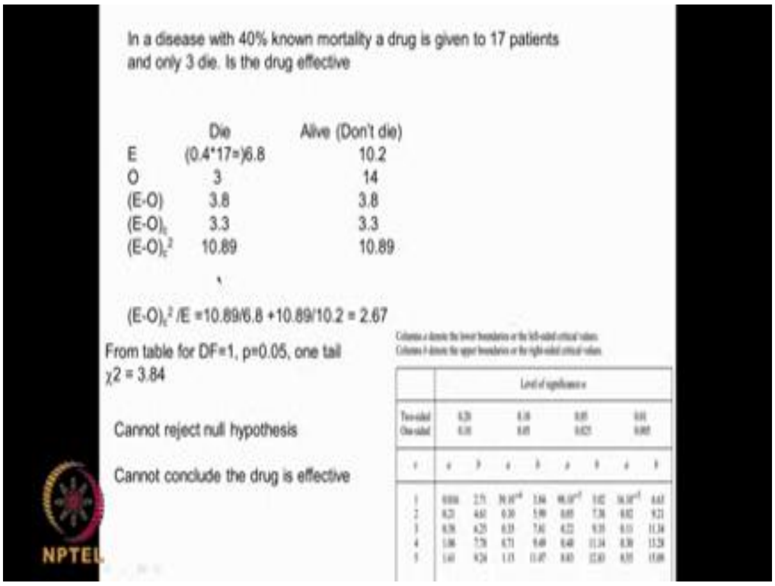
So $[E\text{-}O]_c = [E\text{-}O] - 0.5$

Subtract 0.5 from the absolute of [E-O]
If it becomes <0, then equal it to 0

.

Especially, when you are using only 2 categories, there is something called Yate's correction you need to perform that, especially when you are having a 2 / 2 type of data. That means I have male, female, success, failure. So male, female could be in the column, success, failure could be in the row, that is called a 2 / 2 contingency table. In such situations, it is suggested that we do something called Yate's correction because • χ^2distribution is very continuous. Whereas when we have only 2 by 2 it is almost like an integer, it is almost like a binary you know male, female, success, failure, no, yes sort of things, this especially here. So, what we do is, we take the absolute value of E - O that is expected - O observed, subtract it by 0.5, it is very simple. But if it becomes less than 0, when it subtracts then you will take it as 0 only, do not forget. We take the absolute whether it is E - 1 or O - E, subtract 0.5 from there and then perform the calculation and then go ahead actually that will be the corrected and after that you square it up and then divide it by E - O ÷E and so on actually, ok?. Let us look at a problem, so what you do is, we subtract 0.5 from the absolute value of E - O.

(Refer Slide Time: 10:03)



In a disease with 40% known mortality a drug is given to 17 patients and only 3 die. Is the drug effective

	Die	Alive (Don't die)
E	(0.4*17=)6.8	10.2
O	3	14
(E-O)	3.8	3.8
$(E\text{-}O)_c$	3.3	3.3
$(E\text{-}O)_c^2$	10.89	10.89

$(E\text{-}O)_c^2/E = 10.89/6.8 + 10.89/10.2 = 2.67$

From table for DF=1, p=0.05, one tail

$\chi 2 = 3.84$

Cannot reject null hypothesis

Cannot conclude the drug is effective

Columns a denote the lower boundaries or the left-sided critical values.
Columns b denote the upper boundaries or the right-sided critical values.

	Level of significance α							
Two-sided	0.20		0.10		0.05		0.02	
One-sided	0.10		0.05		0.025		0.01	
ν	a	b	a	b	a	b	a	b
1	0.016	2.71	$3.9 \cdot 10^{-3}$	3.84	$9.8 \cdot 10^{-4}$	5.02	$1.6 \cdot 10^{-4}$	6.63
2	0.21	4.61	0.10	5.99	0.05	7.38	0.02	9.21
3	0.58	6.25	0.35	7.81	0.22	9.35	0.11	11.34
4	1.06	7.78	0.71	9.49	0.48	11.14	0.30	13.28
5	1.61	9.24	1.15	11.07	0.83	12.83	0.55	15.09

We will look at a problem, in a disease with 40 % known mortality. They know that if somebody gets the disease, probability of them dying is 40 %, now a drug is given to 17 patients and only 3 die. So, can we say the drug is effective, that means I am generally if somebody gets the disease 40 % of them will die, 17 patients was given a new drug only 3 died. So, can we say this drug it be effective. Here it is almost like a 2 / 2 data, right?. We have dying, alive. So, 17 people we expect 40 % of 17 to die, that means 0.4 of 17 to die 6.8, so how many will be alive 17 - 6.8 that is 10.2.

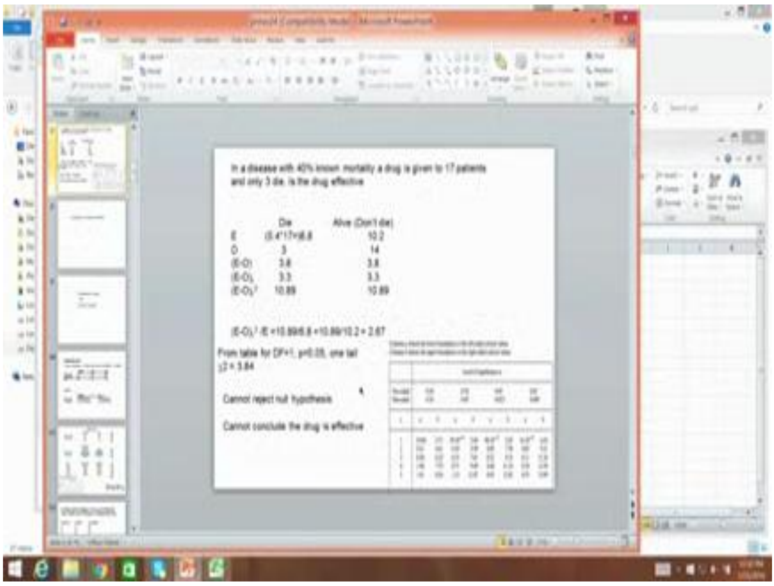
But we observed is 3 died, 14 are alive, expected is 17 * 0.4, 6.8 to die, 10.2 to be alive. So, E - O 3.8, so E - O actually you may get is as negative, it does not matter we put it as 3.8 and so we make a correction here. Subtract 0.5, subtract 0.5 then you square it up then divided, you square it up you get 10. You subtract 0.5 it comes to 3.3, square of 3.3 is 10.89, same thing here 3.3^2 is 10.89. So, 10.89 ÷ expected. 10.89 ÷expected, so we get 2.67. So, for degrees of freedom of 1, 95 %. We see here 3.84 that is your chi square value. We cannot reject the null hypothesis, you understand? We cannot reject the null hypothesis because you get a 3.84 is a table, your statistic test statistics is 2.67. So, cannot reject the null hypothesis. So, cannot conclude the drug is effective actually from this results. We can also do the same calculation using excel or Graphpad also, right?.

(Refer Slide Time: 12:30)

Can determine from EXCEL or GRAPHPAD

If you remember we can do it by excel also or in Graphpad also.

(Refer Slide Time: 12:39)



In a disease with 40% known mortality a drug is given to 17 patients and only 3 die. Is the drug effective

	Die	Alive (Don't die)
E	(0.4*17=6.8	10.2
O	3	14
(E-O)	3.8	3.8
$(E-O)_c$	3.3	3.3
$(E-O)_c^2$	10.89	10.89

$(E-O)_c^2/E = 10.89/6.8 + 10.89/10.2 = 2.67$

From table for DF=1, p=0.05, one tail $\chi2 = 3.84$

Cannot reject null hypothesis

Cannot conclude the drug is effective

.

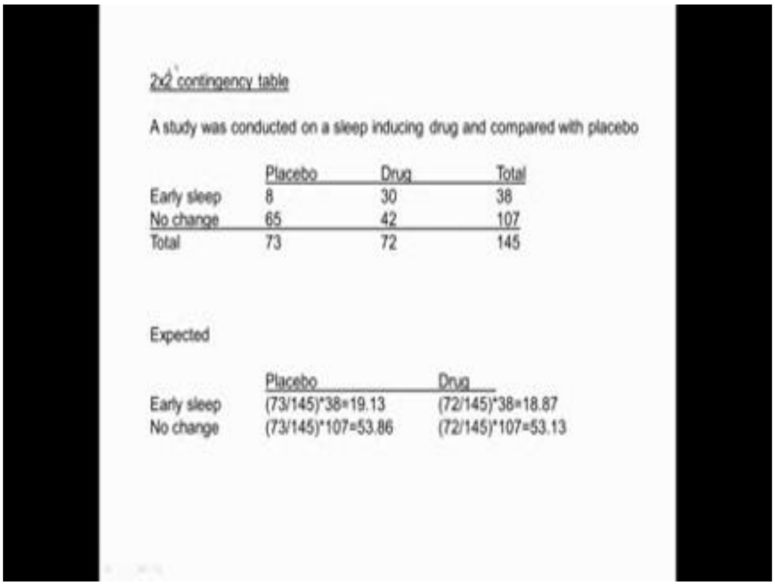
Let us try it do it in excel first. So, we expect here 6.8, whereas 10.2 had a problem then we expect 10.2, so observed is 3, observed is 14 so this is expected. Expected we expect at 6.8 and 10.2 alive whereas observed is this much. Can we do it by, We have the • χ^2 test here. So, we give this range, we give this range. So, it gives you a 0.015, if you give this and this. If we take it as a 95 % two-sided then you get it as 0.05. So, we can also do it through 2 sample, we can also do using the Graphpad also, we can perform this calculation using this type of things. So, the table, the • χ^2 as you know the • χ^2 distribution table.

So, we have the • χ^2 test here, CHITEST, it gives you the range. We give the actual range, the expected range so it calculates and tells you. We can use this particular • χ^2 test also for our calculations. Here we say, we cannot reject the null hypothesis so cannot conclude that drug, drug is effective. Let us look at another problem, of course we can do the same thing using the binomial distribution also. What is the binomial distribution? We have 17 data points and we see 3 of them die so the probability is 0.4 and we can do this particular binomial distribution. And we can say binom, So, 3 of them die, that is 3 observed out of the total data of 17, p is 0.4, then we say either true or false. If we give it as false you get 0.34, if you give it as a true 3,17,0.4, sorry, so 40 %. So, I made it as 0.4 %, sorry, so it comes to 0.46.

According to the binomial distribution we can say that we cannot reject the, I mean we can reject the null hypothesis because the probability comes over to be < 0.05, but when we perform this type of calculation using • χ^2, we end up getting the value much smaller than the table value. The test statistics comes out to be 2.67. So, the main thing is here, we have done a correction for Yate's, if you do not do a Yate's correction, obviously, the test statistics will come out to be much larger because here, we have subtracted 0.5 so the test statics statistics will come out to be much larger if I do not subtract the Yate's correction. Then it may become significant.

So, the question is, many of the softwares might not do the Yate's correction unless, we specifically want to incorporate that actually and ideally we must include Yate's correction especially if the data set is very small like a 2 / 2, this is a 2 / 2 type of a contingency expected out, observed, die, alive that point you need to consider actually. Now let us look at another problem.

(Refer Slide Time: 19:01)



2x2 contingency table

A study was conducted on a sleep inducing drug and compared with placebo

	Placebo	Drug	Total
Early sleep	8	30	38
No change	65	42	107
Total	73	72	145

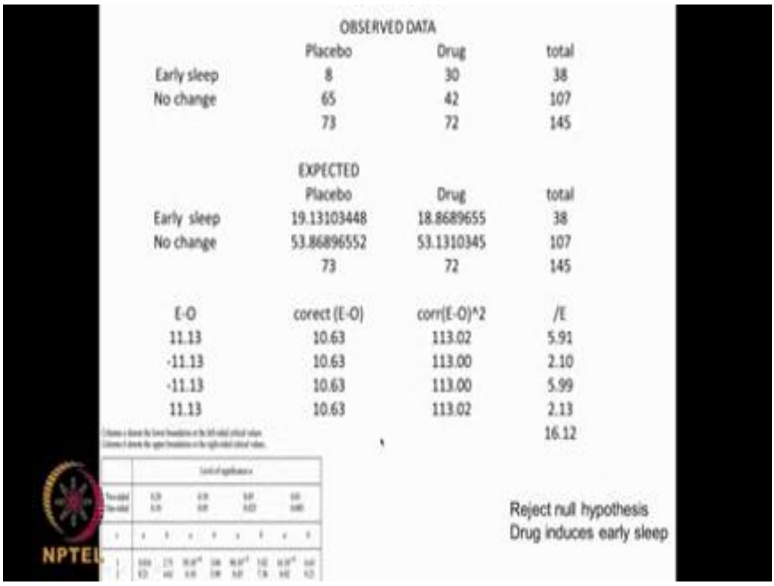
Expected

	Placebo	Drug
Early sleep	(73/145)*38=19.13	(72/145)*38=18.87
No change	(73/145)*107=53.86	(72/145)*107=53.13

This is typical 2 / 2 contingency type of situation. A study was conducted on a sleep inducing drug and it was compared with placebo. Drug was being tested to see whether it induces sleep. So placebo, drug, whoever took placebo, 8 of them said they slept early and 65 of them said there is no change. 30 of the people who took drug said they slept early, 42 of the people who took drug said there is no change. So, we have this type of situation, it was tested on different groups of people as you can see the numbers, do not add up but with the • $\chi2$ distribution does not matter even if that numbers do not add up, that means 2 sets of samples on which it was tested and the placebo was given to 73 people the drug was given to 72 people, right?. Even if the numbers do not add up, we can do still • $\chi2$ test without any problem. In this problem, what we have is 8 of the people who took placebo had a early sleep but 65 of the people who took placebo they said there is no change and 30 of the 72 people who took the drug said they had slept early, 42 of the people who took the drug in that 72 lot said there is no change. We do not know anything about the expected, we know only about what is observed.

So how do we calculate this it is quite simple, so for this one what we do is 73 ÷145 * 38. So, we are doing it the same ratio here 73 ÷ 145. That is the ratio of people who took placebo out of that total, placebo plus drug and in that 38 said early sleep so 73 / 145 *38. That is, the expected-early sleep who took placebo that comes to 19.13. If you want to have the early sleep for drug what you do 72 ÷145 * 38 that gives you 18.87 that is the expected for early sleep-drug. For here what you do 73 ÷145 * 107, 73 ÷145 is the % of 145 people who took placebo and 107 is the set of people who said no change so the expected should be 73 / 145 * 107. And for this 72 / 145 * 107 so that is comes to 53.13. So, we have 4 observed and we calculated 4 expected this is a 2 by 2 contingency table.

(Refer Slide Time: 22:02)



OBSERVED DATA

	Placebo	Drug	total
Early sleep	8	30	38
No change	65	42	107
	73	72	145

EXPECTED

	Placebo	Drug	total
Early sleep	19.13103448	18.8689655	38
No change	53.86896552	53.1310345	107
	73	72	145

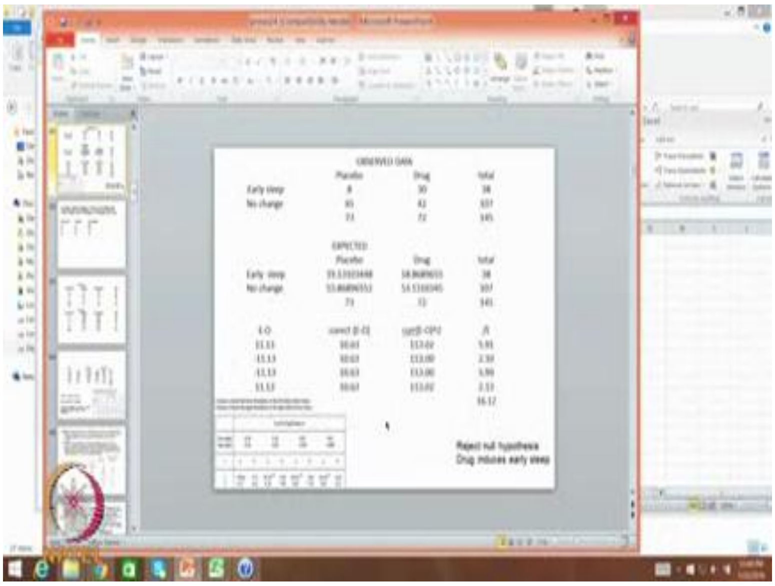
E-O	corect (E-O)	corr(E-O)^2	/E
11.13	10.63	113.02	5.91
-11.13	10.63	113.00	2.10
-11.13	10.63	113.00	5.99
11.13	10.63	113.02	2.13
			16.12

Reject null hypothesis
Drug induces early sleep

.

Let us put it once again in the table form, so from the previous table and then what we calculated here right. Now expected minus observed what do we do here? (E-O) there are 1, 2, 3, 4 so 4 such of data for each one of them expected, (E-O), (E-O), (E-O), (E-O). Now we need to do a correction so 0.5 subtract, as you can see I take the absolute value on subtract 0.5, 0.5 reduction. Then I square it up, that comes to 113 all these then divide by expected, so for each one I divide and then I add up, I get 16.12. With 16.12 is what I get and what is my degrees of freedom placebo drug. So, two categories that means I have 1 degrees of freedom, early sleep no change, again degrees of freedom is 1 so 1 * 1 is 1. So, degrees of freedom here is 1 and for 95 % I will get it as 3.84. The statistics · $\chi2$ is much larger than the · $\chi2$ e of the table. So, what I can say is I can reject the null hypothesis. So, the null hypothesis is there is no difference between the drug and the placebo, but I am rejecting the null hypothesis. So I can say the drug induces early sleep. We can do this problem of course using your excel, using your Graphpad also.

(Refer Slide Time: 24:04)



OBSERVED DATA

	Placebo	Drug	total
Early sleep	8	30	38
No change	65	42	107
	73	72	145

EXPECTED

	Placebo	Drug	total
Early sleep	19.13103448	18.86896551	38
No change	53.86896552	53.13103448	107
	73	72	145

E-O	correct (E-O)	corr(E-O)^2	/E
11.13	10.63	113.02	5.91
11.13	10.63	113.00	2.10
11.13	10.63	113.00	5.99
11.13	10.63	113.02	2.13
			16.12

Reject null hypothesis
Drug induces early sleep

If you go to excel, so what we do? We take these 4 data set. So, I will say (E-O) 19.13, 53.86, 18.86, 53.13 then I go up observed is 8, 65, 30, 42 sorry. 19, 53.86. So, 19, 53.86, 18.86, 53.1, 8, 65, 30 so I can say • $\chi2$ test t e s t, actual that is observed versus expected, observed is this, the expected is this. I get a very small p value so I reject the null hypothesis. So this is the observed, this is the expected in this particular problem. So, we can reject the null hypothesis. We can also do it by the Graphpad software. We have the • $\chi2$ here so continue.

So, we can do statistical distribution interpreting. • $\chi2$ comes here; we can do it by continuous assessment.

So, here we have the • $\chi2$ it give the probability 0.05, degrees of freedom is 1. So, compute • $\chi2$, it gives you 3.84 15. So, for example, if you look at the table here it comes out to be 3.84. So, 16.12 is what you have calculated the test statistics so we can reject the null hypothesis in this particular problem.

We can do using the excel command called the CHITEST or we can use the Graphpad also to calculate the probability or we can use it for calculating the F value. But it is quite simple to do it manually, you can do this very simply by using a simple excel and here please note that we use a something called Yate's correction and when you use an Yate's correction the (E-O) goes down by 0.5. Some softwares might not consider Yate’s correction so you may get slightly different answers. When we do the Yate’s correction you need to take the absolute value and then subtract 0.5.

So here for example, - 11.13 we still subtracted 0.5, we made it absolute and then subtracted. When you do the 0.5 subtraction if the result goes below 0, we should take it as 0, so remember that. It is very simple to do that and once you calculate the test statistics which is

summation of

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

you compare it with the table and then you see whether the table value is greater or less than the test statistics. If the table value is greater then there is no reason for reject the null hypothesis, if the table value is less then you reject the null hypothesis. In this problem as you can see the table value is less than the test statistics, so obviously, we reject the null hypothesis which is drug and placebo are the same status quo. That is your null hypothesis, alternate drug induces early sleep, so we accept the alternate hypothesis.

I also showed you how to do this calculation using excel also, so it is not very difficult for you to do it on excel, the command is CHITEST. So, we will continue on this • χ^2 distribution in the next class also.

Thank you very much.

Key Words: Chi square test, Mean, Variance, Variables, one sided 95%, two sided 95%, Graphpad, binomial distribution, test statistics