

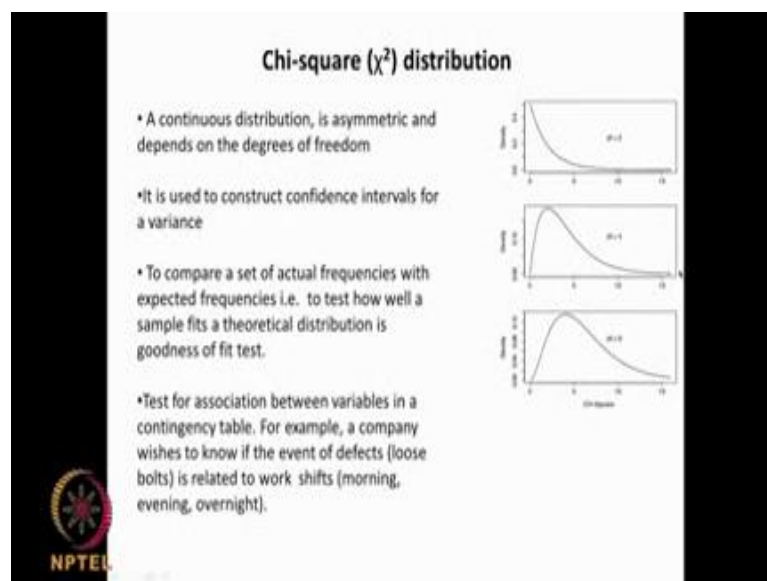
**Biostatistics and Design of Experiments**  
**Prof. Mukesh Doble**  
**Department of Biotechnology**  
**Indian Institute of Technology, Madras**

**Lecture - 23**

**$\chi^2$  distribution/test**

Welcome to the course on Biostatistics and Design of Experiments. We will continue on this  $\chi^2$  distribution as well as  $\chi^2$  test. As I said, it is very important test. It is almost like a binary- yes, no, success, failure type of thing. It is a continuous distribution, it is a skewed distribution and it is very useful if I am comparing observed versus expected, I expect something but I observe something. Is there a statistical difference in this observation, so in such situations I use this  $\chi^2$  test. So, it is distribution once again to recall.

(Refer Slide Time: 00:47)



It is a continuous distribution but it is asymmetric as a degrees of freedom increases. I can see the curve becomes more uniformed but, originally it is right skewed. So, we can construct confidence interval, we can compare actual frequencies with expected frequencies; we can compare fitted data versus the real model, sorry, with the real data. We can look at association between variables. Like for example, I said, is bad **workmanship** related to the work shift, that sort of situations we can look at. Is the poor

results we get because of certain instruments? And that sort of comparative studies can be done using  $\chi^2$  test.

(Refer Slide Time: 01:35)

**Cumulative probability and  $\chi^2$  distribution**

- The total area under the curve is equal to 1.
- The area under the curve between 0 and  $\chi^2$  value is a cumulative probability.

The shaded area is a cumulative probability associated with a  $\chi^2$  statistic equal to 5

It gives the probability that the value of  $\chi^2$  statistic will fall between 0 and 5.

**NPTEL**

So the area under the curve is 1 and so when you are talking about say  $\chi^2$  of 5 then this area will be the cumulative probability that is between 0 and 5 of the occurrence.

(Refer Slide Time: 01:49)

**Critical Values of the  $\chi^2$  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.

Columns A denote the lower boundaries or the left-tailed critical values. Columns B denote the upper boundaries or the right-tailed critical values.

Significance Level	Level of significance $\alpha$				
	0.10	0.05	0.025	0.01	0.005
df	1	2	3	4	5
1	1.645	1.960	2.306	2.706	3.000
2	1.650	1.920	2.278	2.575	2.773
3	1.646	1.890	2.247	2.499	2.669
4	1.645	1.860	2.215	2.423	2.575
5	1.645	1.830	2.183	2.348	2.485
6	1.645	1.810	2.160	2.303	2.433
7	1.645	1.790	2.142	2.271	2.398
8	1.645	1.770	2.127	2.246	2.368
9	1.645	1.760	2.114	2.226	2.346
10	1.645	1.750	2.102	2.209	2.328
11	1.645	1.740	2.091	2.195	2.313
12	1.645	1.730	2.081	2.183	2.300
13	1.645	1.720	2.071	2.172	2.289
14	1.645	1.710	2.062	2.162	2.280
15	1.645	1.700	2.053	2.153	2.272
16	1.645	1.690	2.045	2.145	2.265
17	1.645	1.680	2.037	2.138	2.259
18	1.645	1.670	2.030	2.132	2.254
19	1.645	1.660	2.023	2.126	2.250
20	1.645	1.650	2.017	2.121	2.246
21	1.645	1.640	2.011	2.116	2.242
22	1.645	1.630	2.006	2.112	2.238
23	1.645	1.620	2.001	2.108	2.235
24	1.645	1.610	1.996	2.104	2.232
25	1.645	1.600	1.992	2.101	2.229
26	1.645	1.590	1.988	2.097	2.226
27	1.645	1.580	1.984	2.094	2.223
28	1.645	1.570	1.980	2.091	2.220
29	1.645	1.560	1.976	2.088	2.217
30	1.645	1.550	1.972	2.085	2.214
31	1.645	1.540	1.968	2.082	2.211
32	1.645	1.530	1.964	2.079	2.208
33	1.645	1.520	1.960	2.076	2.205
34	1.645	1.510	1.956	2.073	2.202
35	1.645	1.500	1.952	2.070	2.199
36	1.645	1.490	1.948	2.067	2.196
37	1.645	1.480	1.944	2.064	2.193
38	1.645	1.470	1.940	2.061	2.190
39	1.645	1.460	1.936	2.058	2.187
40	1.645	1.450	1.932	2.055	2.184
41	1.645	1.440	1.928	2.052	2.181
42	1.645	1.430	1.924	2.049	2.178
43	1.645	1.420	1.920	2.046	2.175
44	1.645	1.410	1.916	2.043	2.172
45	1.645	1.400	1.912	2.040	2.169
46	1.645	1.390	1.908	2.037	2.166
47	1.645	1.380	1.904	2.034	2.163
48	1.645	1.370	1.900	2.031	2.160
49	1.645	1.360	1.896	2.028	2.157
50	1.645	1.350	1.892	2.025	2.154

**NPTEL**

Now, there is a table just like your t table, F table, z table and so on. We have  $\chi^2$  table also, in this column we have the degrees of freedom, here we have for the two sided and

here the one sided. This gives you the lower boundary on the left hand side of the critical value because this curve is like this, right? so you will a lower boundary and upper boundary. Generally we look on that side, upper side. So this is a two sided, this is a one sided. So obviously, two sided 0.05 is one sided 0.025, two sided 0.1 is one sided 0.05. If the test statistics is greater than the upper critical value, then we reject the null hypothesis or if it is lower than this lower critical value also, then we reject the null hypothesis.

(Refer Slide Time: 02:47)

**$\chi^2$ -test for a population variance**

To determine the difference between a sample variance  $s^2$  and a population variance  $\sigma_0^2$ .

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

To test the null hypothesis that the population variance is equal to  $\sigma_0^2$ .

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.

NPTEL

We can first do a test for seeing whether the sample comes from a population, the sample variance is coming from the population variance. So if I have a sample set  $x_1, x_2$  up to  $n$  I calculate a variance  $s^2$ , then I can compare it with the population variance of  $\sigma_0^2$ . The null hypothesis they are both same, the sample is same from the population or alternate will be they are not in the same population. So the test statistic is like this

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

,  $n$  is the data points,  $s^2$  is the variance of the sample,  $\sigma_0^2$  is the variance of the population.

(Refer Slide Time: 03:27)

The  $\chi^2$ -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

$H_0 : O_i = E_i$   
 $H_1 : O_i \neq E_i$

The test statistic is compared with the critical value from  $\chi^2$  tables with  $\nu$  DF. Where,  $\nu = k - 1$ .

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



Then, we can use this for testing goodness of fit. So, I observe something but I expect something,

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

, add up. So here the null hypothesis

$$H_0 : O_i = E_i$$

$$H_1 : O_i \neq E_i$$

, alternate will be . Again we compare it, the test statistics with the table and then if the test statistics is greater than the table critical value, we will say the null hypothesis can be rejected.



expected, right?.

So you get a  $\chi^2$  of 0.625. Now if you look, go to your table 0.05 how many degrees of freedom? You have 2 machines so 1, we have 3 grades 3 that gives you 2, so totally  $2 * 1$  is 1 sorry, 2. So we look in the under the column of 2 and then we say 5.99 for a 95 %. So, we expect the table is 5.99, test is 0.625 so null hypothesis cannot be rejected. So the grades are independent of the machine, very nice problem. Let us look at another situation.

(Refer Slide Time: 06:14)

Incidence of three types of malaria in three tropical regions.


	Asia	Africa	South America	Totals
Malaria A	31	14	45	90
Malaria B	2	5	53	60
Malaria C	53	45	2	100
Totals	86	64	100	250

Ho: The two categorical variables are independent. (no relationship between location and type of malaria)  
•H1: The two categorical variables are related.

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} = \chi^2 = 125.516$$

Degrees of Freedom =  $(c - 1)(r - 1) = 2(2) = 4$   
Critical value: 9.488 ( $\alpha = 0.05$ )

Would reject the null hypothesis that there is no relationship between location and type of malaria



We have say incidence of 3 types of Malaria and 3 tropical regions you know Asia, Africa, South America. Malaria is a very serious problem, there are different types of malaria- malaria A, malaria B, malaria C.

So, the malaria A is happening 31, 14, 45, malaria B is 2, 5, 53, malaria C is 53, 45, 2, right?. So, I want to know whether it follows the expected, this is what is observed. The expected, we expect all the malaria either to be same, that means the **Ho** will be the 2 categorical variable that is type of malaria versus the type of continents should be independent, there should not be any relationship on that actually. The other one is the alternate will be these two variables are related to each other. So if there are 90 cases, we expect that all of them should be there in all the countries because, we do not expect any sort of correlation or relationship in this entire situation.

So this is the observed, expected is equally probable in all the 3 continents Malaria A, Malaria B, Malaria C should be equally probable that will be what is called expected. So observed is this, so

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

so we should be able to get the test statistics. In this particular problem, we get 125.5, so this problem we get it as 125.5 And degrees of freedom we have 3 continents, so 2 is the degree, 3 types of malaria, 2 is degrees of freedom, for type, so  $2 * 2$  is 4. So we go to 4 degrees of freedom and 95 %. So let us look at the table, 4 degrees of freedom and 95 % so we get 9.49, so 9.488. So when you do that, we will get test statistics is 125.5, the table is 9.488, so we can reject the null hypothesis. What is the null hypothesis? There is no relationship between the type of malaria and the continent location, so we can reject. So there seems to be some sort of a correlation, on type of malaria and the type of and the location or place from which it is observed. This is how we need to do, this type of problem of where we have the observed verses expected. Now as I said Excel also can do, there are two types of functions in Excel.

(Refer Slide Time: 09:28)

The screenshot shows an Excel spreadsheet with the following content:

**Chi-Square Test for Independence**

These different grades of product is inspected from two machines.

Grade	M1	M2	Total
A	10	10	20
B	10	10	20
Total	20	20	40

Null Hypothesis:  $H_0$ : The two categorical variables are independent.  
 Alternative Hypothesis:  $H_1$ : The two categorical variables are related.

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

$\chi^2 = 0.025$

From table,  $\chi^2_{0.05} = 3.84$

$0.025 < 3.84$

$\chi^2_{obs} < \chi^2_{crit}$

Conclusion: Hypothesis valid. The grades are independent of the machines.

I mentioned one is the CHITEST, where in CHITEST we give the actual value then we give the expected value. So let us look at these previous problem of grade where we say

we expect, observed is 3, 9, 8, 7, 11, 12 whereas expected is 4, 8, 8, 6, 12, 12. So how do you do this problem, CHITEST. So the actual expected, actual and the expected. So we get a probability of 0.98 so obviously, null hypothesis cannot be rejected. Do you understand? So that is what we got from our calculation also or we can say CHIINV 0.05, the degrees of freedom here is 3 gives you 2, 2 machines gives you 1, that gives 2 so it is 5.991. So we need to look at this 5.99 here actually, right?. So the CHI INVERSE gives you exactly like your table, it gives you the  $\chi^2$  value and that is the critical value.

Whereas, the CHITEST gives you the probability for your problem and gives you the probability, given the observed and the expected, we can do, now let us look at this problem also using excel and see whether we can do it using excel. So this is the observed, **sorry**, observed is 31, 2, 53, 14, 5, 45, 45, 53, 2 whereas when I say expected. How do I calculate the expected? It is bit tricky to do that actually, we look at that is this type of problem later on actually, we need to calculate expected is  $86 \div 250 * 90$ . So it will be like  $86 \div 250 * 90$ . There is we expect 30.96 because, 86 is the summation of this, 250 is your grand total, the summation of this is 90. So you take that same ratio you will get it as  $86 \div 250 * 90$ . So for this, what is that ratio?  $86 \div 250 * 60$ ,  $86 \div 250 * 60$ . What about this one? This will be  $86 \div 250$ ,  $86 \div 250 * 100$ , that is 34.4 that means, I expect 34.4, I expect 20.64, I expect 30.96. Do you understand how to get this each one of them? So if I want to get this I say,  $86 \div 250$  that is the grand total multiplied by 90. So it is just the question of ratio.

Say, if I want to get this 53, I will say  $86 \div 250 * 100$ . Then if I want to get this the expected what I do,  $64 \div 250 * 90$ . The next one will be  $64 \div 250$ ,  $\div 250 * 60$ . The next one will be  $64 \div 250 * 100$ . The next one here will be for an equivalent 45 so what do I do,  $100 \div 250 * 90$ . Then next one will be  $100 \div 250$ ,  $100 \div 250 * 60$ . Then next one will be  $100 \div 250$ ,  $100 \div 250 * 100$ , understand?. So that is how we got, sorry, 100. There is this minus term here I made a mistake,  $100 \div 250 * 100$ . So we can cross check by doing summation it come to 250 and this also should come to, there is a mistake there somewhere so obviously, this is also should come out to be 250 so it is not come to be 250 so obviously, we made a mistake. So the first term here will be  $86 \div 250 * 90$  and so on actually.

$86 \div 250 * 90$ , then next one will be  $86 \div 250 * 60$ , then  $86 \div 250 * 100$ , then next one will be  $64 \div 250 * 90$ , then  $64 \div 250 * 60$ , then  $64 \div 250 * 100$ , so obviously, there is a



mistake I have made here so I put minus here so that is the mistake. So I got 250 in both the cases that is very good. So I can use chi test, **CHITEST comma** so you get a very small probability which means that, we can reject the null hypothesis, which means that null hypothesis there is no relationship between the location or continent and the type of malaria. Same thing I got here right, I rejected.

Now, so by using a CHITEST, I am able to reject the null hypothesis. Now same thing we can do by CHI INVERSE also, I will say 0.05 the degrees of freedom as you know I said is 4 because 3 continents and 3 types of disease so 4 so it comes out to be 9.48 right?, so 9.48. So from the CHITEST command, we can reject the null hypothesis. We can also do it using your Graphpad because of, as you can see we have the commands for Graphpad here also, we have the **x<sup>2</sup>** here, right? **chi square** or here we have the **chi square** which can do this type of calculations. So we can do **chi square** exactly, so **chi square** from a probability this is equivalent to a chi inverse whereas this equivalent to a CHITEST command. So you can see this CHI TEST command.

So it is exactly like the CHITEST command where it gives you the p value or in this particular case, it will give you, the given the probability it will give the chi value it is exactly like chi inverse. So even in your Graphpad we can do that so the CHITEST and the CHIINV command available in excel, even in Graphpad we have use these two we use this to do that actually. So let us go back to our problem and so we will say we will reject the null hypothesis that, there is an association between these two categories or parameters.

(Refer Slide Time: 19:11)



So one important point you need to consider is generally, if you are involving only two categories for example, if I am doing a **2 by 2** contingency table especially for two then there is something called Yate's correction because  $\chi^2$  distribution is continuous whereas when we use two categories-yes, no or success, failure or drug working, not working obviously, we are not able to as such do good justice, because in a two category system it is like a discontinuous data or integer data whereas  $\chi^2$  is a continuous data. So in such situations we need to subtract something called the Yate's correction. So this is called a Yate's correction, where you just do the  **$E_i - O_i$  observed, that is you calculate  $E_i - O_i$  then subtract minus 0.5 from there.**

After that you square it up and then divide by, that after that it is also same expected, no problem. So if you look at the normal, we used to have  **$O_i - E_i$** <sup>2</sup> or  **$E_i - O_i$** <sup>2</sup>, it divided by expected. So what you do is, you just subtract **minus 0.5** this is valid only when you have something like a **2 by 2** type of contingency table and then in such situations we just, instead of this term you subtract from, **minus 0.5** from there and then after that you square it up and divide expected in the denominator, so that is all. That is only difference which you need to consider when you are going to give Yate's correction this is valid when we are talking about **2 by 2** type of contingency table, where we are talking in terms of integers especially drug working, not working, yes, no, not and yes that sort of situations we need to use it.

So we will look at some problems where we use Yate's correction, but one important point is when the data is negative that is when you calculate **expected minus observed** negative we cannot further subtract from **minus 0.5** into that.

(Refer Slide Time: 21:29)

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	$(0.4 \cdot 17) = 6.8$	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  
 $\chi^2 = 3.84$

Cannot reject null hypothesis  
 Cannot conclude the drug is effective

NPTEL

		Level of significance			
Two-tail	One-tail	0.20	0.10	0.05	0.01
Two-tail	One-tail	0.10	0.05	0.025	0.005
1	0.000	2.71	3.84	3.84	6.63
2	0.20	4.01	5.02	5.02	7.38
3	0.25	4.35	5.39	5.39	7.78
4	0.30	4.61	5.62	5.62	8.00
5	0.35	4.78	5.79	5.79	8.15
6	0.40	4.90	5.90	5.90	8.25
7	0.45	5.00	6.00	6.00	8.34
8	0.50	5.09	6.09	6.09	8.43
9	0.55	5.18	6.18	6.18	8.51
10	0.60	5.26	6.26	6.26	8.59
11	0.65	5.34	6.34	6.34	8.66
12	0.70	5.41	6.41	6.41	8.73
13	0.75	5.48	6.48	6.48	8.79
14	0.80	5.54	6.54	6.54	8.85
15	0.85	5.60	6.60	6.60	8.91
16	0.90	5.66	6.66	6.66	8.96
17	0.95	5.71	6.71	6.71	9.01
18	1.00	5.76	6.76	6.76	9.06
19	1.05	5.81	6.81	6.81	9.11
20	1.10	5.86	6.86	6.86	9.16
21	1.15	5.91	6.91	6.91	9.21
22	1.20	5.96	6.96	6.96	9.26
23	1.25	6.01	7.01	7.01	9.31
24	1.30	6.06	7.06	7.06	9.36
25	1.35	6.11	7.11	7.11	9.41
26	1.40	6.16	7.16	7.16	9.46
27	1.45	6.21	7.21	7.21	9.51
28	1.50	6.26	7.26	7.26	9.56
29	1.55	6.31	7.31	7.31	9.61
30	1.60	6.36	7.36	7.36	9.66
31	1.65	6.41	7.41	7.41	9.71
32	1.70	6.46	7.46	7.46	9.76
33	1.75	6.51	7.51	7.51	9.81
34	1.80	6.56	7.56	7.56	9.86
35	1.85	6.61	7.61	7.61	9.91
36	1.90	6.66	7.66	7.66	9.96
37	1.95	6.71	7.71	7.71	10.01
38	2.00	6.76	7.76	7.76	10.06
39	2.05	6.81	7.81	7.81	10.11
40	2.10	6.86	7.86	7.86	10.16
41	2.15	6.91	7.91	7.91	10.21
42	2.20	6.96	7.96	7.96	10.26
43	2.25	7.01	8.01	8.01	10.31
44	2.30	7.06	8.06	8.06	10.36
45	2.35	7.11	8.11	8.11	10.41
46	2.40	7.16	8.16	8.16	10.46
47	2.45	7.21	8.21	8.21	10.51
48	2.50	7.26	8.26	8.26	10.56
49	2.55	7.31	8.31	8.31	10.61
50	2.60	7.36	8.36	8.36	10.66

That is very, very important. When we are talking in terms of Yate's correction that means, when **E-O** is already negative we cannot again do more negative and take it out below 0, that is not permitted actually. Let us look at problem, which is **2 by 2** sort of situation because we are talking about a drug working on a set of patients so because of the drug, patients either die or alive, we expect some data whereas we observe something else so we need to look at whether there is an association. So in such situation we are also applying the Yate's correction as well, in this particular problem.

Look at this particular problem. In a disease with 40 % known mortality, a drug is given to 17 patients and only 3 die. **So 40 %** are known to die that means, 17 patients we expect 6.8 to die, the remaining 10.2 not to die. But we observe 3 to be dying and 14 to be alive, right? **3 + 14 is 17**, 6.8. So is the drug effective?, we need to find out. What we do? **E-O** that is **E - O, 6.8 minus 3 is 3.8 (E-O)**.

Then we do a correction here, and then that is we subtract it by 0.5 that comes to 3.3. Then, once you calculate the difference between the expected and the observed, you use the absolute value and then subtract **minus 0.5** from there, that is the Yate's correction. And that is why you get the corrected **E minus O**, that is **expected minus observed** to be

the same on both sides, because we are taking the absolute value and then subtracting **minus 0.5**. After that you square them and then you divide each one of them with the expected, you add up to get **expected minus observed** corrected square by expected. Now this is what you get 2.67. Now this is at 1 degree of freedom because we have die, alive, 2 states so we have 1 degree of freedom. So if you look under 1 degree of freedom for a **p = 0.05** for a 1 sided test, we get the table **chi square** as 3.84. Now your statistics is 2.67 and table is 3.84 so you cannot reject the null hypothesis. So we cannot conclude the drug is effective. So you understand this Yate's correction here, so remember that you do the Yate's correction that is you subtract **minus 0.5** from the absolute value of **expected minus observed**.

Thank you very much.

**Key Words: Confidence interval, frequencies, variables, cumulative probability, degree of freedom, critical value, upper boundary, lower boundary, the expected, the observed, 2x2 type of contingency table, absolute value, Yate's correction.**