

**Research Methodology**  
**Prof. Soumitro Banerjee**  
**Department of Physical Sciences**  
**Indian Institute of Science Education and Research, Kolkata**

**Lecture - 46**  
**Hypothesis Testing: The Chi - Square Test - Part 02**

(Refer Slide Time: 00:17)

$E \neq 0$   
 $(E - O)^2$   
 Random variable  $Y$   
 $(E_i - O_i)^2$   
 $\chi^2 = \sum_i \frac{(E_i - O_i)^2}{E_i}$   
 $Q_1 = Y_1^2 \dots k=1$   
 $Q_2 = Y_1^2 + Y_2^2 \dots k=2$   
 $Q_3 = Y_1^2 + Y_2^2 + Y_3^2 \dots$   
 $\vdots$   
 $Q_n = Y_1^2 + Y_2^2 + \dots + Y_n^2$

$\mu = 0$   
 $\sigma = 1$   
 $\chi^2$  table  
 Level of significance  
 = (1 - Level of confidence)

$\chi^2 = 3.841$   
 5%

$freq(O_i)$   
 $k=1$   
 $k=2$   
 $k=3$   
 $k=4$

So, let me illustrate how this is actually done.

(Refer Slide Time: 00:24)

$\chi^2$  test  
 the value of some quantity A should be  $E$ . Observed value  $O$ .

$E \neq 0$   
 $(E - O)^2$   
 Random variable  $Y$   
 $(E_i - O_i)^2$   
 $\chi^2 = \sum_i \frac{(E_i - O_i)^2}{E_i}$   
 $Q = Y_1^2 \dots k=1$

$\mu = 0$   
 $\sigma = 1$   
 $\chi^2$  table  
 Level of significance  
 = (1 - Level of confidence)

$\chi^2 = 3.841$   
 5%

$freq(O_i)$   
 $k=1$   
 $k=2$   
 $k=3$   
 $k=4$

Now, there are situations where we try to test how certain elements belong to categories; sometimes we test for categories.

(Refer Slide Time: 00:44)

The slide content is as follows:

Chi<sup>2</sup> test for categories

$E_i$  --- expected number in bin  
 $O_i$  --- observed " " " bin

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

The slide also features the NPTEL logo in the top right corner and a small inset video of a speaker in the bottom right corner.

So, chi square test for categories. What do you mean by categories? Categories are sort of 'bins' in which you can put elements. For example the classical Mendel experiment, in which he was crossing a tall variety of pea plants and a short variety of pea plants, and in the first generation all the plants were tall. In the second generation, when these plants were crossed with each other, he found that some are tall, some are short. The theory of genetics later predicted that it should be 3:1 ratio in which the tall variety and the short variety will appear.

So, in this case a particular plant is either tall or short; which means that, it can be put in either this bin or that bin; put into this category or that category. And a theory predicts certain number of elements or experimental subjects to be in a particular bin. It predicts another particular number to be in another particular bin. Then we apply the chi square test for these categories.

So, in that case what are our expectation values? E is the expected number in a bin, and O is the observed number in that bin. Since there can be different categories, this will have to be subscripted by i. So, one can put in one category and then it would be  $E_1$  and  $O_1$ , another category  $E_2$ ,  $O_2$  and so on and so forth. So, the theory will predict that

so many will be in bin 1, so many will be in bin 2, so many will be in bin 3 and so on and so forth.

And the test will allow you to check the correctness of that kind of a theory. But in that case, how do we define the chi square? As we said, it is nothing but the E minus O. The order does not really matter, In some books you will find O minus E, sometimes you will find E minus O. It does not matter because you are ultimately squaring it.

So, it is  $O_i$  minus  $E_i$  square and you need to normalize it. Otherwise you will not get the proper sense of how big is the error compared to what is the expected value. That is why you need to normalize it; and it is summed over  $i$ .

Now, if you define this way, you are essentially saying that I have obtained a value and my expected value was this. Therefore, this is the fluctuation, this is the difference between the observed value and the expected value, when normalized. That means, we are essentially getting a sense of how big are the differences between the expectation and the observed values.

If the theory is true, then these differences should be due to completely random fluctuations. And therefore, this whole thing should behave like random variables, and if there is only one  $i$ ; that means, only one of these, then we will have to tally the behaviour of only one random variable. If there are two of them we will have to tally it with a combination of two random variables; that means,  $Y_1$  square plus  $Y_2$  square, its character.

$$\begin{array}{ll}
 Q_1 = Y_1^2 & k = 1 \\
 Q_2 = Y_1^2 + Y_2^2 & k = 2 \\
 Q_3 = Y_1^2 + Y_2^2 + Y_3^2 & k = 3 \\
 \vdots & \vdots \\
 Q_n = Y_1^2 + Y_2^2 + Y_3^2 + \dots + Y_n^2 & k = n
 \end{array}$$

If there are three such bins, three such categories, then we have to tally it with 3 of them:  $Y_1$  square plus  $Y_2$  square plus  $Y_3$  square, how does that behave. And these things are available in the chi square tables. Let me illustrate with an example.

(Refer Slide Time: 06:13)

Sample: Tall/short, round/wrinkled

Independently assorted

Tall + round + short + wrinkled

Tall + round

9:3:3:1

Type	Expected	Observed
Tall, round	90	84
Tall, wrinkled	30	34
Short, round	30	38
Short, wrinkled	10	4
Total	160	160



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

$$= \frac{(84 - 90)^2}{90} + \frac{(34 - 30)^2}{30} + \frac{(38 - 30)^2}{30} + \frac{(4 - 10)^2}{10}$$

$$= 6.46$$

DoF = n - 1

The critical value of  $\chi^2$  for rejecting the hypo. = 7.815

Now, let us go to an example. You know there are certain characteristics in plants. For example, let us take the example of the pea plant in which there are two categories: tall or short, that is regarding the height of the plant. And there is also another category: whether the peas are round or wrinkled. That means, the peas themselves are roundish or these are corrugated, wrinkled.

So, these two characters we find and we do not know whether these two characters are related in some way or other. But, the theory of genetics says that if these two characters are 'independently assorted', then, if you cross a tall variety with round seed with short plus wrinkled seeds, then in the first generation they are all tall plus round seeds. But, in the second generation there are four possibilities: tall and round, tall and wrinkled, short and round, short and wrinkled. All of them appear, but in a specific ratio 9:3:3:1.

So, if they are independently assorted, then they should appear in this ratio. Now, suppose somebody wants to test whether two such characters are truly independently assorted or not. And suppose she has conducted the experiment. She has chosen tall and round, and short and wrinkled plants, and crossed them. In the second generation she finds that there are all four possible varieties of plants. She counts them. She picks up some number of plants and then she counts them, sorts them and puts them into categories.

And suppose the result is something like this, that you have put in form of a table. Here we have the type and there are four different types: tall, round, tall, wrinkled, short, round, and short wrinkled. Now, here we need to have some expectation and we need to have some observation. So, let me allow two columns.

Types	Expected (E)	Observed (O)	$(O - E)^2 / E$
Tall, round	90	84	$(-6)^2 / 90$
Tall, wrinkled	30	34	$4^2 / 30$
short, round	30	38	$8^2 / 30$
short, wrinkled	10	4	$(-6)^2 / 10$
Total	160	160	

Here the expected and here it is the observed. Now, what is expected? What is expected is the ratio should be 9:3:3:1. But you actually do not observe the ratio. You observe the numbers. So, suppose the total is 160 plants that she counted, and she observed 84 to be tall and round, 34 to be tall and wrinkled, 38 to be short and round and only 4 to be short and wrinkled. That is what she found and the summation is 160.

Now, the question is, what is expected? What is expected is this: if you add this up 9 plus 3 plus 3 plus 1 turns out to be 16; so, 9 out of 16 should be tall and round. So, you can easily see that 9 by 16 into 160 is equal to 90. So, basically whatever is the number you have to multiply by that by 10. So, what is expected is 90 numbers should be tall and round, 30 numbers should be tall and wrinkled; so, these are the numbers that are expected ok.

Now, on that basis, we can go ahead and do the calculation. This also adds up to 160. So, you see that there is a difference between the observation and the expectation values as we expected from the theory, but in spite of this difference can we still say that the theory is correct that these two characters are independently assorted? This is the question that we are asking.

So, in that case we have to calculate the chi square. The chi square will be, as we have already seen, it is basically the observation  $i$  minus expectation  $i$  square by expectation of the  $i$ th value. So, this will be for the first one 90 minus  $i$  do not really care whether this

one should come first. But, let me write it that same way; so, that you are not confused 84 minus 90 square divided by 90 plus 34 minus 30 square divided by 30 plus 38 minus 30 square divided by 30 plus 4 minus 10 square divided by 10.

$$\begin{aligned}\chi^2 &= \sum_i \frac{(O_i - E_i)^2}{E_i} \\ &= \frac{(-6)^2}{90} + \frac{4^2}{30} + \frac{8^2}{30} + \frac{(-6)^2}{10} \\ &= 6.66\end{aligned}$$

So, 84 is observed minus 90 expected square divided by the expected value. We have calculated each one of them, each category, and then we have summed them up and this number comes to be 6.66. Now, the question is that, how do we use this?

For that we need to refer to the table. I will show how to read the table. But, the point is that we are trying to figure out whether they are independently assorted, we are trying to make the assertion with 95 percent confidence and therefore, significance level of 0.05. So, significance level is 1 minus the confidence level and therefore, we are trying to figure it out with a significance level of 0.05. We will now have to read from the table.

But, what is the degree of freedom? Notice there are four of these categories. So, common sense would say that the degree of freedom is 4; you have to consider 4 of them. But no. The point is that if you know the total number that you had started with, the total number that you picked up and counted. Then if you know three of them, you can calculate the 4th by adding this and subtracting from this one.

Therefore, once you have obtained these three, the fourth one does not have a freedom. There is no freedom to choose the fourth one. Therefore, the degree of freedom is actually n minus 1. So, in this case it is 4 minus 1; so, 3 degrees of freedom.

(Refer Slide Time: 17:11)

df	$\chi^2_{.995}$	$\chi^2_{.990}$	$\chi^2_{.975}$	$\chi^2_{.950}$	$\chi^2_{.900}$	$\chi^2_{.800}$	$\chi^2_{.700}$	$\chi^2_{.500}$	$\chi^2_{.300}$	$\chi^2_{.100}$
1	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188

Now, let us see how we read from the table. Now, here you can see the chi square table and for any value of the degree of freedom these are tabulated here, and for each degree of freedom you have got a row. And the area beyond a certain value of chi square is tabulated. What is tabulated is the area beyond a certain value of chi square.

So, for example, we are now looking for the significance level of 0.05 which is chi square, the subscript is 0.05. It is this column that we are looking for. And we are looking for a row corresponding to the degree of freedom of 3. So, this row and this column and here we have 7.815. So, the chi square value is 7.815 beyond which the area is 5 percent of the total area; that means, 0.05 significance level; so, this is the value that we are looking for.

And if we have a chi square value beyond this, then we would say that it is very unlikely to have that chi square value coming out of purely random variables. While if it is somewhere here, below that critical value of chi square, then we would say that we do not have sufficient ground for rejecting the null hypothesis.

The table says that, the 0.05 level of significance happens at 7.815. So, the critical value of chi square for rejecting the hypothesis, in this case the null hypothesis, comes out to be 7.815. I will show you how actually it is done. So, this is the value that you get from the chi square table and the value that you actually got from the experiment is 6.66,

which is lower. Lower means that the probability of getting this value is more than 5 percent, which means that we cannot reject the null hypothesis.

(Refer Slide Time: 20:37)

↓  
9:3:3:1

Short, Criminal	10	4
Total = 160	160	160

NPTEL

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

$$= \frac{(84 - 70)^2}{70} + \frac{(34 - 30)^2}{30} + \frac{(38 - 30)^2}{30} + \frac{(4 - 10)^2}{10}$$

$$= 6.66$$

DoF = 1 - 1

The critical value of  $\chi^2$  for rejecting the hypo. = 7.815

Null is true. ∴

$H_0 =$  Categories are independently assorted  
 $H_1 =$  " " not " "

In this example, I have not written the hypothesis and the null hypothesis. Let me do that otherwise it will be a bit confusing. The null hypothesis is that the categories are independently assorted, and the alternative hypothesis categories are not independently assorted. So, if they are independently assorted then these would be the expected values. That is what we have written. If they are not independently assorted, then it will be a different value.

So, we are always checking for the equalities and we would be able to reject the null hypothesis if the value of chi square is beyond this value, but it is not. And therefore, we have to conclude that the null is true; there is not enough basis to reject the null hypothesis.