Chi – Square Distribution

The sum of squares of n standard normal variables is defined to be a chi – square variable with *n* d.f.

$$\chi^2 = X_1^2 + X_2^2 + \ldots + X_n^2 .$$

The χ^2 variable can be used to perform a variety of tests.

Uses of χ^2 test / distribution

1. **To test the independence of two or more attributes**:

   χ^2 test can be used to determine whether there is any association between two or more attributes (characteristics) like colour of eyes of mothers and daughters, heights of fathers and sons etc.

   In these tests, we proceed on the null hypothesis that the attributes are independent i.e. there is *no association* between the attributes.

TEST OF INDEPENDENCE OF ATTRIBUTES – CONTINGENCY TABLES

Consider two attributes A & B. Let A be divided into r classes A_1, A_2,., A_r and B be divided into s classes B_1 ,B_2,B_3,….,,B_s

[For instance attribute A can be colour of hair; A_1:black ; A_2:brown ; A_3 : red ; attribute b can be region; B_1 : Asia ; B_2 : Africa ; B_3 : Europe]

	A	A_1	A_2	A_3	…	A_r	Total
B							
B_1		(A_1B_1)	(A_2B_1)	(A_3B_1)	…	(A_rB_1)	(B_1)
B_2		(A_1B_2)	(A_2B_2)	(A_3B_2)	…	(A_rB_2)	(B_2)
B_3		(A_1B_3)	(A_2B_3)	(A_3B_3)	…	(A_rB_3)	(B_3)
B_4		(A_1B_4)	(A_2B_4)	(A_3B_4)	…	(A_rB_4)	(B_4)
⋮		⋮	⋮	⋮	⋮	⋮	⋮
B_s		(A_1B_s)	(A_2B_s)	(A_3B_s)		(A_rB_s)	(B_s)
Total		(A_1)	(A_2)	(A_3)		(A_r)	N

Such a classification in which attributes are divided into more than two classes is known as *manifold classification*.

The various cell frequencies can be expressed in a table known as *r x s* manifold contingency table where (A_i) is the number of persons possessing attribute A_i, I = 1,2,…,r and (B_j) is the number of persons possessing attribute B_j, j = 1,2,…,s

Also $\sum_{i=1}^{r} A_i = \sum_{j=1}^{s} B_j = N$ where N is the total frequency.

Now, the expected number of persons possessing both the attributes A_iB_j is given by:

(under H_0) $Expected\ frequency = \frac{Row\ total \times column\ total}{overall\ total\ (N)}$

[The RHS can be obtained from the given table]

For large N,

$$\chi^2_{cal} = \sum_i \sum_j \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2_{(r-1)(s-1)}\, d.f$$

Where O_{ij} and E_{ij} are the observed and expected frequencies respectively in the (i,j)th cell. (r and s are the number of rows and columns in the table respectively.

If χ^2_{cal} > $\chi^2_{\text{(r-1)(s-1)}}$ at α *los*, then we reject H_0 and conclude that the attributes are not independent. Otherwise we have no reason to reject H_0. We obtain $\chi^2_{\text{(r-1)(s-1)}}$ at α *los* from the chi-squared table, and it is obtained in the same manner as reading the t distribution. All the areas given in the table are the right tail. Since the question is about the attributes being dependent or independent, we don't have to worry about one-tailed and two-tailed tests.

While testing the independence of attributes, the Null Hypothesis H_0 will always be that the attributes are independent. The Alternative Hypothesis will suggest that they are dependent, i.e. there is some relationship between them.

Prof. Anushri Tambe

Example: Two sample polls of votes for two candidates A and B for a public office are taken, one from among the rural areas and the other from urban areas. The results are given in the following table:

Area	Votes for	
	A	B
Rural	620	380
Urban	550	450

Examine whether the nature of the area is related to voting preference in this election.

Solution: Let H_0 : Area is independent of voting preference.

Consider

Area	Votes for		
	A	B	Total
Rural	620	380	1000
Urban	550	450	1000
Total	1170	830	2000

Under H_0, the expected frequencies are obtained as follows:

Using $Expected\ frequency = \frac{Row\ total \times column\ total}{overall\ total\ (N)}$

Area	Votes for	
	A	B
Rural	$\frac{1170X1000}{2000}$ = 585	$\frac{830X1000}{2000}$ = 415
Urban	$\frac{1170X1000}{2000}$ = 585	$\frac{830X1000}{2000}$ = 415

(That is if the voting preferences were independent of the area that the electorate lived in, then the 1170 votes that A received would have equally come from the rural and urban areas (585 each.)

Calculations for χ^2

Class	Frequency		$o_i - e_i$	$(o_i - e_i)^2$	$\frac{(o_i - e_i)^2}{e_i}$
	o_i	e_i			
Rural A	620	585	35	1225	2.094
Rural B	380	415	35	1225	2.952
Urban A	550	585	35	1225	2.094
Urban B	450	415	35	1225	2.952
Total					10.092

$\chi^2_{cal} = 10.092$; Here r = 2 ; s = 2 ; (r-1)(s-1) = 1

From the tables, $\chi^2_{0.05}$ for 1 d.f. = 3.841

Since $\chi^2_{cal} > \chi^2_{1,0.05}$,

We reject H_0 at 5% *los* and conclude that the nature of area is not independent of voting preference. Hence, the voting preference depends on the area the electorate lives in.

Exercises:

1. Out of 8,000 graduates in a town, 800 are females; out of 1,600 graduate employees, 120 are females. Test if any distinction is made in appointment on the basis of sex.

2. A random sample of students of DBIT was selected and asked their opinions about ‘autonomous colleges’. The results are given below. At 5% l.o.s., test the hypothesis that the opinions of the students are independent of the branch of engineering.

Class	Favouring ‘autonomous colleges’	Opposed to ‘autonomous colleges	Total
EXTC	120	80	200
Mechanical	130	70	200
I.T.	70	30	100
Computers	80	20	100
Total	400	200	600