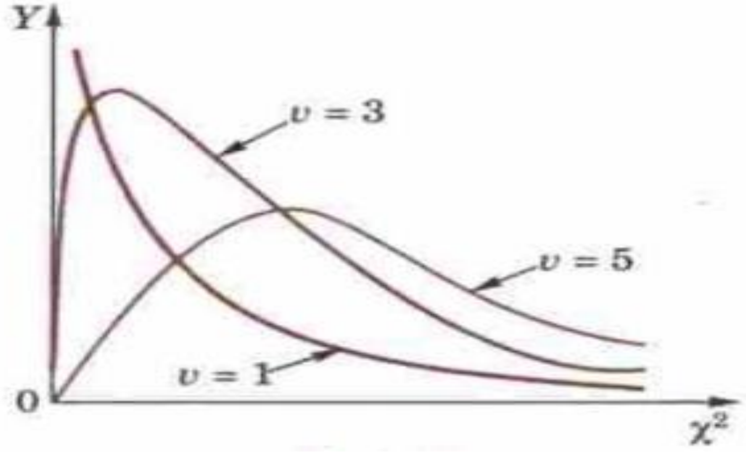
Chi-square test (Non-parametric test)

Chi-square is a measure of actual divergence of the observed and expected frequencies



For degree of freedom>1, this curve is tangential to x-axis at the origin and is Positively skewed

Goodness of fit-

The value of χ^2 is used to test whether the deviations of the observed frequencies from the expected frequencies are significant or not. It is also used to test how will a set of observations fit a given distribution, therefore χ^2 provides a test of goodness of fit and may be used to examine the validity of some hypothesis about an observed frequency distribution. As a test of goodness of fit it can be used to study the correspondence between theory and fact

Procedure to test significance and goodness of fit

(i) Set up null hypothesis

(ii) Calculate $\chi^2 = \sum \frac{\{O-E\}^2}{E}$ where O is observed frequency and E is expected frequency

(iii) Find degree of freedom and ${\chi^2}_{tab}$ from the table.

(iv) If ${\chi^2}_{cal} < {\chi^2}_{tab}$ H_o is accepted otherwise rejected.

Chi-square test

Methods to find Expected Frequency

- Taking average ($\nu = n - 1$)
- Using Binomial or Poisson Distribution ($\nu = n - 2$)
- Using R*C Table (($\nu = (r - 1) * (c - 1)$), $E_{ij} = \frac{A_i * B_j}{N}$

where r:no of rows, c:no of columns

Remarks

- Frequency of every cell should be >=5
- If observations in the given list are < 5 , we group the data with the neighbouring frequency and degree of freedom is reduced accordingly.
- Yates Correction- In a 2*2 table $\nu = (r - 1) * (c - 1)$=1. If any of the cell frequency is <5, we use formula of Chi-square test with Yates Correction.

$$\chi^2 = \sum \frac{\{|O-E|-0.5\}^2}{E}$$

NOTE: Before calculating value of χ^2, value of each cell should be observed so that cells may be merged if required.

Ex 1- The following table the number of aircraft accidents that occur during various days of the week. Find whether the accidents are uniformly distributed over the week.

Days	Sun.	Mon.	Tues.	Wed.	Thurs.	Fri.	Sat.
No. of accidents	14	16	8	12	11	9	14

Solution H_o: the accidents are uniformly distributed over the week.
H_1: the accidents are not uniformly distributed over the week. Table for

Total no of accidents = 84

expected frequencies for each day can be calculated as an average. So E for all O is given as E= 84/7 = 12

$$\chi^2 = \sum \frac{\{O - E\}^2}{E} = \frac{(14-12)^2}{12} + \frac{(16-12)^2}{12} + \frac{(8-12)^2}{12} + \frac{(12-12)^2}{12} + \frac{(11-12)^2}{12} + \frac{(9-12)^2}{12} + \frac{(14-12)^2}{12}$$

= 4.17

Deg of freedom=7-1=6; χ^2 =12.59 at 5% LOS

${\chi^2}_{cal} < {\chi^2}_{tab} \Rightarrow H_o$ is accepted i.e. the accidents are uniformly distributed over the week.

Ex 2- Fit a Poisson distribution to the following data and test the goodness of fit

x	0	1	2	3	4	5	6
f	275	72	30	7	5	2	1

Solution.

Mean, $m = \frac{\Sigma fx}{\Sigma f} = \frac{0+72+60+21+20+10+6}{392} = \frac{189}{392} = 0.482$

Poissondistribution

$$P(r) = \frac{e^{-m} m^r}{r\,!} \quad \text{or} \quad P(r) = \frac{e^{-0.482}\,(0.482)^r}{r\,!}$$

$$P(0) = e^{-0.482} = 0.6175, \quad f(0) = 392 \times 0.6175 = 242.1$$

$$P(1) = \frac{e^{-0.482}(0.482)^1}{1\,!} = 0.2976, \quad f(1) = 392 \times 0.2976 = 116.7$$

$$P(2) = \frac{e^{-0.482}(0.482)^2}{2\,!} = 0.0717, \quad f(2) = 392 \times 0.0717 = 28.1$$

$$P(3) = \frac{e^{-0.482}(0.482)^3}{3\,!} = 0.0115, \quad f(3) = 392 \times 0.0115 = 4.5$$

$$P(4) = \frac{e^{-0.482}(0.482)^4}{4\,!} = 0.00139, \quad f(4) = 392 \times 0.00139 = 0.5$$

$$P(5) = \frac{e^{-0.482}(0.482)^5}{5\,!} = 0.0001, \quad f(5) = 392 \times 0.0001 = 0.1$$

$$P(6) = \frac{e^{-0.482}(0.482)^6}{6\,!} = 0.00001, \quad f(6) = 392 \times 0.00001 = 0$$

$$\chi^2 = \frac{(275-242.1)^2}{242.1} + \frac{(72-116.7)^2}{116.7} + \frac{(30-28.1)^2}{28.1} + \frac{[(7+5+2+1)-(4.5+0.5+0.1)]^2}{4.5+0.5+0.1}$$

$\Rightarrow \chi^2_{cal} = 40.938$

deg of freedom=n-2=4-2;

χ^2 =5.99 at 5% LOS

$\chi^2_{cal} > \chi^2_{tab}$

$\Rightarrow H_o$ is rejected

i.e. Poisson distribution is not good fit.

Ex3 - From the following table, showing the number of plants having certain character, test the hypothesis that the flower colour is independent of flatness of leaf.

	Flat leaves	*Curled leaves*	*Total*
White Flowers	99	36	135
Red Flowers	20	5	25
Total	119	41	160

Solution H_o: flower colour is independent of flatness of leaf.

H_1: flower colour is not independent of flatness of leaf.

Table for expected frequencies $E_{ij} = \frac{A_i * B_j}{N}$ is given below:

	Flat leaves	*Curled leaves*	*Total*
White flowers	$\frac{135 \times 119}{160} = 100$	$\frac{135 \times 41}{160} = 35$	135
Red flowers	$\frac{25 \times 119}{160} = 19$	$\frac{25 \times 41}{160} = 6$	25
Total	119	41	160

$$\chi^2 = \sum \frac{\{O - E\}^2}{E} = \frac{(99-100)^2}{100} + \frac{(36-35)^2}{35} + \frac{(20-19)^2}{19} + \frac{(5-6)^2}{6} = \frac{1}{100} + \frac{1}{35} + \frac{1}{19} + \frac{1}{6} = 0.2579$$

Deg of freedom= (r-1)(c-1)=(2-1)(2-1)=1 ; $\chi^2 = 3.84$at 5% LOS

${\chi^2}_{cal} < {\chi^2}_{tab} \Rightarrow H_o$is accepted i.e.. flower colour is independent of flatness of leaf.

Exercise

Of a group of patients who complained they did not sleep well, some were given sleeping pills while others were given sugar pills (although they all thought they were getting sleeping pills). They were later asked whether the pills helped them or not. The result of their responses are shown in the table given below. Assuming that all patients told the truth, test the hypothesis that there is no difference between sleeping pills and sugar pills at a significance level of 0.05.

	Slept well	*Did not sleep well*
Took sleeping pills	44	10
Took sugar pills	81	35

Ex 4 A set of 5 similar coins is tossed 320 times. Test the hypothesis that the data given below follow a binomial distribution.

No. of heads	*0*	*1*	*2*	*3*	*4*	*5*
Frequency	*6*	*27*	*72*	*112*	*71*	*32*

No. of heads	0	1	2	3	4	5
Frequency	6	27	72	112	71	32

Solution. $P\text{ (Head)} = \frac{1}{2},\ \ q = 1 - \frac{1}{2} = \frac{1}{2}$

Theoretical frequencies are

$$P(0H) = q^5 = \left(\frac{1}{2}\right)^5 = \frac{1}{32}, \text{ Frequency of 0 head} = \frac{320}{32} = 10$$

$$P(1H) = {}^5C_1\ pq^4 = {}^5C_1\left(\frac{1}{2}\right)\left(\frac{1}{2}\right) = \frac{5}{32}, \text{ Frequency of 1 head} = \frac{5}{32}\times 320 = 50$$

$$P(2H) = {}^5C_2\ p^2\,q^3 = 10\left(\frac{1}{2}\right)^2\left(\frac{1}{2}\right)^3 = \frac{10}{32}, \text{ Frequency of 2 heads} = \frac{10}{32}\times 320 = 100$$

$$P(3H) = {}^5C_3\ p^3\,q^2 = 10\left(\frac{1}{2}\right)^3\left(\frac{1}{2}\right)^2 = \frac{10}{32}, \text{ Frequency of 3 heads} = \frac{10}{32}\times 320 = 100$$

$$P(4H) = {}^5C_4\ p^4\,q = 5\left(\frac{1}{2}\right)^4\left(\frac{1}{2}\right) = \frac{5}{32}, \text{ Frequency of 4 heads} = \frac{5}{32}\times 320 = 50$$

$$P(5H) = {}^5C_5\ p^5\,q^0 = \left(\frac{1}{2}\right)^5 = \frac{1}{32}, \text{ Frequency of 5 heads} = \frac{1}{32}\times 320 = 10$$

$$\chi^2 = \sum \frac{\{O-E\}^2}{E} = 78.68$$

Deg of freedom=6-1=5;

$\chi^2 = 11.07$ at 5% LOS

$\chi^2{}_{\text{cal}} > \chi^2{}_{\text{tab}}$

$\Rightarrow H_o$ is rejected

i.e. the data given does not follow a binomial distribution

Ex 5- The theory predicts the proportion of beans in the four groups A, B, C, D should be 9:3:3:1. In an experiment among 1600 beans, the numbers in the four groups were 882, 313, 287 and 118. Does the experimental result support the theory?

Solution H_o: The experimental result supports the theory

H_1: The experimental result does not supports the theory

Total no of beans = 882+313+287+118=1600

These are to be divided to be in the ratio 9:3:3:1. So expected can be written as

$$E(882) = \frac{9}{16} \times 1600 = 900, \quad E(313) = \frac{3}{16} \times 1600 = 300$$

$$E(287) = \frac{3}{16} \times 1600 = 300, \quad E(118) = \frac{1}{16} \times 1600 = 100$$

$\chi^2 = \sum \frac{\{O-E\}^2}{E}$= 4.7266 ;

Deg of freedom=4-1=3

$\chi^2 = 7.815$ at 5% LOS

${\chi^2}_{cal} < {\chi^2}_{tab}$

$\Rightarrow H_o$ is accepted

i.e. The experimental result supports the theory.

Exercise

1. Genetic theory states that children having one parent of blood type M and other blood type N will always be one of three types M, MN, N and that the proportions of these types will on average be 1 : 2 : 1. A report states that out of 300 children having one M parent and one N parent, 30% were found to be of type M, 45% of type MN and remainder of type N. Test the hypothesis by x^2 test.

2. A bird watching sitting in a park has spotted a number of birds belonging to 6 categories. The exact classification is given below:

Category	1	2	3	4	5	6
Frequency	6	7	13	17	6	5

Test at 5% level of significance whether or not the data is compatible with the assumption that this particular park is visited by birds belonging to these six categories in the proportion

= 1 : 1 : 2 : 3 : 1 : 1.

Exercise

1.

Fit a Binomial Distribution to the data

x	0	1	2	3	4	5
f	38	144	342	287	164	25

and test for goodness of fit at the level of significance 0.05.

2.

Two hundred digits were chosen at random from a set of tables. The frequencies of the digits were as follows:

Digits	0	1	2	3	4	5	6	7	8	9
Frequency	18	19	23	21	16	25	22	20	21	15

Use x^2 test to assess the corrections of hypothesis that the digits were distributed in equal numbers in the table from which they were chosen.

3.

The following is the distribution of the hourly number of trucks arriving at a company's warehouse.

Trucks arriving per hour :	0	1	2	3	4	5	6	7	8
Frequency :	52	151	130	102	45	12	5	1	2

Find the mean of the distribution and using its mean, (rounded to one decimal) as the parameter λ, fit a Poisson distribution. Test for goodness of fit at the level of significance $\alpha = 0{\cdot}05$

Ex 6- Two batches of 12 animals each are given test of inoculation. One batch was inoculated and the other was not . The number of dead and surviving animals are given in the following table for both cases. Can the inoculation be regarded as effective against the disease at 5% LOS.

	Dead	Surviving	Total
Inoculated	2	10	12
Not-inoculated	8	4	12
Total	10	14	24

Solution $H_o There\ is\ no\ association\ between\ inaculation\ and\ death.$

$H_1: There\ is\ \ association\ between\ inaculation\ and\ deat$ Ther expected frequencies $E_{ij} = \frac{A_i * B_j}{N}$ are given below:

E(2)=5, E(10)=7, E(8)=5, E(4)=7

Deg of freedom= (r-1)(c-1)=(2-1)(2-1)=1 ; Here data can not be merged as in that case degree of freedom will become zero, So we use Yate's correction formula.

$\chi^2 = \sum \frac{\{|O-E|-0.5\}^2}{E}$ = 4.29

χ^2 =3.81 at 5% LOS

${\chi^2}_{cal} > {\chi^2}_{tab} \Rightarrow H_o$is rejected i.e.. inoculation is effective against the disease at 5% LOS.