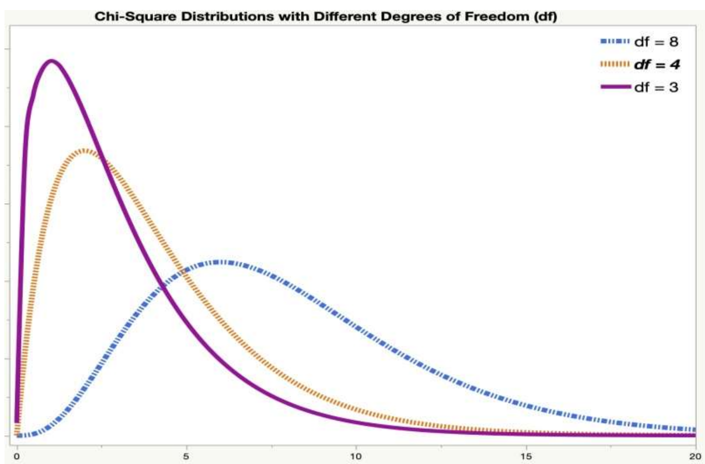
Chi – square (χ^2) Test

Chi-square variate: The square of a standard normal variate is known as chi-square (χ^2) variate with one degree of freedom.

If $x_1, x_2, x_3, \ldots, x_n$ be n independent normal variates with mean zero and S.D unity, then it can be shown that $x_1^2 + x_2^2 + \ldots x_n^2$, is a random variate having χ^2 - distribution with n d.f.

The equation of the χ^2 - curve is

$$y = y_0 e^{-\frac{\chi^2}{2}} (\chi^2)^{\frac{\nu - 1}{2}} \quad \text{where } \nu = n-1$$



Properties of χ^2 - distribution:

1. The sum of independent chi-square variables is also a chi-square variable.
2. χ^2 distribution tends to normal as $n \to \infty$.
3. The number of independent variables is usually called the number of degrees of freedom denoted by ν.

4. If k is the number of linearly independent constraints then ν=n-k
5. For p x q contingency table, degrees of freedom is (p – 1) (q – 1)

Conditions for validity of χ^2 test:

This is an approximate test for large values of n,. For the validity of chi-square test of ‘goodness of fit’ between theory and experiment , the following conditions must be satisfied.

1. The sample observations should be independent.
2. Constraints on the cell frequencies, if any, must be linear,($\sum O_i = \sum E_i$)
3. N, the total frequency, should be reasonably large, greaterthan 50.
4. No theoretical cell frequency should be less than 5. If any theoretical frequency is less than 5, then for the application of χ^2 test, it is pooled with the succeeding or preceding so that the pooled frequency is more than 5.

Application or uses of χ^2 distribution:

1. To test the goodness of fit.
2. To test the independence of attributes.
3. To test if the hypothetical values of the population variance is σ^2.
4. To test the homogeneity of independent estimates of the population variance.

Chi-square Test of goodness of fit:

Definition: If $O_1, O_2, O_3, \ldots O_n$ be a set of observed frequencies and $E_1, E_2, E_3, \ldots E_n$ be the corresponding set of expected (theoretical frequencies, then χ^2 is defined by the relation

$$\chi^2 = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} + \frac{(O_3 - E_3)^2}{E_3} + \ldots + \frac{(O_n - E_n)^2}{E_n}$$

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

with n-1 degrees of freedom.

Independence of attributes :

To test the independent of attributes we calculate χ^2 value from the cell frequencies and compare with table values of χ^2 corresponding to degrees of freedom. For a 2 x2 contingency table

a	b
c	d

Then the χ^2 value is $\chi^2 = \frac{(a+b+c+d)(ad-bc)^2}{(a+b)(a+c)(b+d)(c+d)}$

Degree of freedom is (2-1) (2-1)=1

In a 2 x 2 table, if the frequencies of a cell is small, say 3 or 4, we make Yate's correction to make χ^2 continuous.

Yate's correction Calculate **ad** and **bc ,** and add 0.5 to both factors of the small product and subtract 0.5 to both factors of larger product.

After Yates's correction:

$$\chi^2 = \frac{N\left(ad - bc - \frac{N}{2}\right)^2}{(a+b)(a+c)(b+d)(c+d)}, \qquad N = Total\ \ frequency$$

Procedure to test significance of attributes and goodness of fit:

1. Set up a null hypothesis and calculate

$$\chi^2 = \sum_{i=1}^{n} \frac{\left(O_i - E_i\right)^2}{E_i}$$

2. Find the degrees of freedom and the read the corresponding values of χ^2 at a prescribed significance level from Table.
3. From χ^2 table , we can also find the probability p=0.05 corresponding to the calculated values of χ^2 for the given d.f.
4. If P < 0.05, the observed value of χ^2 is significant at 5 % level of significance.
   If P < 0.01, then the value of significant is 1% level.
   On the other hand if P> 0.05 then it is not significant.

1. In experiments on pea breeding, the following frequencies of seeds were obtained.

Round and yellow	Wrinkled and yellow	Round and green	Wrinkled and green	Total
315	101	108	32	556

Theory predicts that the frequencies should be in proportion 9:3:3:1. Examine the correspondence between theory and experiment.

H_0: There is no difference between theoretical and experimental frequency.

H_1: There is difference between theoretical and experimental frequency.

As the theoretical frequency is given as proportion we have

Round and yellow	Wrinkled and yellow	Round and green	Wrinkled and green	Total
$\frac{9}{16}\times 556 = 313$	$\frac{3}{16}\times 556 = 104$	$\frac{3}{16}\times 556 = 104$	$\frac{1}{16}\times 556 = 35$	556

Hence $\chi^2 = \sum_{i=1}^{n} \frac{\left(O_i - E_i\right)^2}{E_i}$

i.e.,

$$\chi^2 = \frac{(315-313)^2}{313} + \frac{(101-104)^2}{104} + \frac{(108-104)^2}{104} + \frac{(32-35)^2}{35}$$

$$= \frac{4}{313} + \frac{9}{104} + \frac{16}{104} + \frac{9}{35} = 0.51$$

For ν=4 – 1 = 3 d.f., $\chi_{0.05}^{2} = 7.815$

Since the calculated of χ^2 < table value of χ^2 . Accept the null hypothesis ,

There is no difference between theoretical and experimental frequency.

Chi-square test to test INDEPENDENCE OF ATTRIBUTES.

Let us consider two attributes A and B divided in r classes$A_1, A_2, \ldots A_r$ and B divided into s classes $B_1, B_2, B_3, \ldots Bs$. If this is expressed as r x s matrix , then the matrix is called r x s contingency table.

Note : For a 2 x 2 contingency table

a	b
c	d

$$\chi^2 = \frac{(a+b+c+d)(ad-bc)^2}{(a+b)(a+c)(b+d)(c+d)}$$ can be used to save time

Observed frequency	Attribute B_1	Attribute B_2
Attribute A_1	a	b
Attribute A_2	c	d

Expected frequency $= A_i B_j = \dfrac{(A_i)(B_j)}{N}$

That is expected frequency in each cell is

$$= \frac{\text{Product of column total and row total}}{\text{whole total}}$$

2. In an investigation into the health and nutrition of two groups of children of different social status, the following results are got

Social status / Healths	Poor	Rich	Total
Below normal	130	20	150
Normal	102	108	210
Above normal	24	96	120
Total	256	224	480

Discuss the relationship between the Health and their social status.

H_0: There is relationship between Health and social status.

H_1: There is no relationship between Health and social status.

Social status / Healths	Poor	Rich	Total
Below normal	$\frac{256\times150}{480}=80$	$\frac{150\times224}{480}=70$	150
Normal	$\frac{256\times210}{480}=112$	$\frac{2224\times210}{480}=98$	210
Above normal	$\frac{256\times120}{480}=64$	$\frac{224\times120}{480}=56$	120
Total	256	224	480

$$\chi^2=\frac{(130-80)^2}{80}+\frac{(20-70)^2}{70}+\frac{(102-112)^2}{112}+\frac{(108-98)^2}{98}+\frac{(24-64)^2}{64}+\frac{(96-56)^2}{56}$$

$$=31.25+35.71+0.89+1.02+25+28.57=122.44$$

d.f = (3 – 1) x(2 – 1) = 2

$\chi_{0.05}^{2}=5.991$ -- Table value

χ^2 calculated value is more than the table value . Reject the null hypothesis.

Therefore there is no relationship between social status and health.

Note 1: In case of

Fitting a Binomial distribution degrees of freedom = n – 1 .

Fitting a Poisson distribution degrees of freedom = n – 2 .

Fitting a Normal distribution degrees of freedom = n – 3 .

Note 2: If χ^2=0, all observed and theoretical (expected) frequencies coincide.