CHI-SQUARE DISTRIBUTION

The χ^2 distribution was first obtained by Helmert in 1875 and rediscovered by Karl Pearson in 1900.

The square of a standard normal variate is known as chi-square variate with 1 degree of freedom (d.f).

Thus if $X` \sim N(\mu, \sigma^2)$, then $Z = \frac{X-\mu}{\sigma} \sim N(0,1)$ and $Z^2 = \left(\frac{X-\mu}{\sigma}\right)^2$, is a chi-square variate with 1 d.f abbreviated by the letter χ^2 of the Greek alphabet.

In general, if $X_1, X_2, ..., X_n$ are n independent normal variates with means $\mu_1, \mu_2, ..., \mu_n$ and standard deviations $\sigma_1, \sigma_2, ..., \sigma_n$ respectively then the variate

$$\chi^2 = \left(\frac{X_1-\mu_1}{\sigma_1}\right)^2 + \left(\frac{X_2-\mu_2}{\sigma_2}\right)^2 + ... + \left(\frac{X_n-\mu_n}{\sigma_n}\right)^2$$

$$= \sum_{i=1}^{n}\left(\frac{X_i-\mu_i}{\sigma_i}\right)^2$$

Which is the sum of squares of n independent standard normal variates, follows chi-square distribution with n d.f.

APPLICATIONS OF CHI-SQUARE DISTRIBUTION

χ^2 distribution has a large number of applications, some of which are listed below:

1. Chi-square test of goodness of fit.
2. Chi-square test for independence of attributes.
3. Chi-square test for the population variance.

1. CHI-SQUARE TEST OF GOODNESS OF FIT

A very powerful test to describe the magnitude of discrepancy between theory and observation was given by Prof. Karl Pearson in 1900. It enables us to find if the deviations of the observations from theory is just by chance or is it really due to the inadequacy of the theory to fit the observed data. This test is known as χ^2-test of goodness of fit.

If $O_i(i = 1,2,...,n)$ is a set of observed frequencies and $E_i(i = 1,2,...,n)$ is the corresponding set of expected (theoretical) frequencies, then the Statistic χ^2 may be defined as

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

follows chi-square distribution with (n-1) d.f.

In order to determine whether the divergence is due to chance or otherwise. We have to compare the computed value of χ^2 with the table values. Table values of χ^2 as given by R.A. Fisher are available for various levels of confidence, ordinarily up to 30 degrees of freedom. If the calculated value of χ^2 is less than the table value at the particular level of confidence, the divergence is said to arise due to fluctuations of sampling. If the calculated value of χ^2 exceeds the table value, the divergence is said to be significant.

Illustration: A die is thrown 132 times with the following results:

Number turned up: 1 2 3 4 5 6

Frequency: 16 20 25 14 29 28

Test the hypothesis that the die is unbiased.

Solution: Null Hypothesis: Set up the null hypothesis that the die is unbiased.

On the basis of hypothesis that the die is unbiased, we expect each number to turn up, 132/6=22 times

Apply χ^2-test

O	E	$(O-E)^2$	$\frac{(O-E)^2}{E}$
16	22	36	1.64
20	22	4	0.18
25	22	9	0.41
14	22	64	2.91
29	22	49	2.23
28	22	36	1.64
			$\sum \frac{(O-E)^2}{E}$ =9.01

No of degrees of freedom = n-1=6-1=5

For 5 degrees of freedom at 5% level of significance, the table value of χ^2=11.07. The calculated value of χ^2 is less than the table value and hence there is no evidence against the hypothesis that die is unbiased.

Illustration: The theory predicts the proportion of beans in the four groups A, B, C and 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 theory?

Solution: Null Hypothesis: We set up the null hypothesis that the experimental results support the theory.

On the basis of hypothesis, the theoretical frequencies can be computed as follows:

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

These can be divided in the ratio 9:3:3:1

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

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

Apply χ^2-test

O	E	$(O-E)^2$	$\frac{(O-E)^2}{E}$
882	900	324	0.3600
313	300	169	0.5633
287	300	169	0.5633
118	100	324	3.2400
			$\sum\frac{(O-E)^2}{E}$ =4.7266

No. of degrees of freedom=n-1=4-1=3

For 3 d.f. at 5% level of significance, the table value of χ^2=7.815. The calculated value of χ^2 is less than the table value. Hence the null hypothesis may be accepted at 5% level of significance and conclude that there is good correspondence between theory and experiment.

2. CHI-SQUARE TEST FOR INDEPENDENCE OF ATTRIBUTES

Under this test, we can find out whether two or more attributes are associated or not. Let us consider two attributes A and B, A is divided into 'r' classes $A_1, A_2,..., A_r$ and B is divided into 's' classes $B_1, B_2,..., B_s$. 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 the following table known as r×s manifold contingency table where (A_i) denotes the number of persons possessing the attribute $A_i, (i = 1,2,..., r)$, (B_j) denote the number of persons possessing the attribute $B_j, (j = 1,2,..., s)$ and (A_iB_j) denote the number of persons possessing both the attributes (A_i) and (B_j). Also $\sum_{i=1}^{r} A_i = \sum_{j=1}^{s} B_j = N$, is the total frequency.

r×s contingency table

A / B	A_1	A_2	....	A_i	....	A_r	Total
B_1	(A_1B_1)	(A_2B_1)	....	(A_iB_1)	....	(A_rB_1)	(B_1)
B_2	(A_1B_2)	(A_2B_2)	....	(A_iB_2)	....	(A_rB_2)	(B_2)
⋮		⋮		⋮		⋮	⋮
B_j	(A_1B_j)	(A_2B_j)	....	(A_iB_j)	....	(A_rB_j)	(B_j)
⋮		⋮		⋮		⋮	⋮
B_s	(A_1B_s)	(A_2B_s)	....	(A_iB_s)	....	(A_rB_s)	(B_s)
Total	(A_1)	(A_2)	....	(A_i)	....	(A_r)	N

Under the null hypothesis that the two attributes A and B are independent, the expected frequencies are calculated as follows

$P(A_i)$ = probability that a person possessing the attribute A_i

$$= \frac{(A_i)}{N}; \quad i = 1,2,..., r$$

$$= \frac{(B_j)}{N}; \quad j = 1,2,..., s$$

$P(A_iB_j) = P(A_i)P(B_j)$ (attributes A_i and B_j are independent under the null hypothesis)

$$P(A_iB_j) = \frac{(A_i)}{N} \cdot \frac{(B_j)}{N}$$

If $(A_iB_j)_o$ denote the expected frequency of (A_iB_j), then

$$\left(A_iB_j\right)_o = N.P\left(A_iB_j\right) = \frac{\left(A_iB_j\right)}{N}, \quad (i = 1,2,...,r\,;\ j = 1,2,...,s)$$

By using this formula expected frequencies for each of the cell frequencies $\left(A_iB_j\right), \ (i = 1,2,...,r\,;\ j = 1,2,...,s)$ can be worked out.

The exact test for independence of attributes is very complicated but a fair degree of approximation is given, for large samples by the χ^2-test of goodness of fit i.e

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{s}\left[\frac{\left(\left(A_iB_j\right)-\left(A_iB_j\right)_o\right)^2}{\left(A_iB_j\right)_o}\right]$$

follows χ^2 distribution with (r-1)(s-1) degrees of freedom.

Now comparing this calculated value with the tabulated value for (r-1)(s-1) d.f at certain level of significance, we reject or retain the null hypothesis of independence of attributes at that level of significance.

Illustration: A certain drug was administered to 456 males out of total 720 in a certain locality to test its efficacy against typhoid. The incidence of typhoid is shown below. Find out the effectiveness of the drug against the disease

	Infection	No infection	Total
Administering the drug:	144	312	456
Without administering the drug:	192	72	264
Total:	336	384	720

Solution: We set up the null hypothesis that the two attributes "incidence of typhoid" and the administration of the drug are independent.

Under the hypothesis of independence.

$$E(144) = \frac{336\times 456}{720} = 212.8$$

$$E(312) = \frac{384\times 456}{720} = 243.2$$

$$E(192) = \frac{336\times 264}{720} = 123.2$$

$$E(72)=\frac{264\times 384}{720}=140.8$$

Apply χ^2-test

O	E	$(O-E)^2$	$\frac{(O-E)^2}{E}$
144	212.8	4733.44	22.244
192	123.2	4733.44	38.420
312	243.2	4733.44	15.171
72	140.8	4733.44	65.742
			$\sum\frac{(O-E)^2}{E}$ =141.577

Degrees of freedom=(r-1)(s-1)=(2-1)(2-1)=1 d.f

For 1 d. f at 5% level of significance the table value of χ^2=3.84. Since calculated value is very much greater than the table value. It is highly significant. Hence the null hypothesis is rejected at 5% level of significance and we conclude that the drug is certainly effective in controlling typhoid.

3. CHI-SQUARE TEST FOR THE POPULATION VARIANCE

Suppose we want to test if the given normal population has a specified variance $\sigma^2=\sigma_o^2$ (say).

Under the null hypothesis that the population variance $\sigma^2=\sigma_o^2$, the statistic

$$\chi^2=\sum_{i=1}^{n}\left[\frac{(x_i-\bar{x})^2}{\sigma_o^2}\right]=\frac{1}{\sigma_o^2}\left[\sum_{i=1}^{n}x_i^2-\frac{\sum x_i^2}{n}\right]=\frac{ns^2}{\sigma_o^2}$$

follows chi-square distribution with (n-1) d.f.

Comparing calculated value with tabulated value of χ^2 for (n-1) d.f at certain level of significance, we may retain or reject the null hypothesis.

Illustration: A random sample of size 10 from a normal population gave the following values:

65, 72, 68, 74, 77, 61, 63, 69, 73, 71

Test the hypothesis that population variance is 32.

Solution: We set up the null hypothesis $H_o : \sigma^2 = 32$ against the alternative $H_1 : \sigma^2 > 32$.

Computation of sample variance

x	$(x-\bar{x})$	$(x-\bar{x})^2$
65	-4.3	18.49
72	2.7	7.29
68	-1.3	1.69
74	4.7	22.09
77	7.7	59.29
61	-8.3	68.89
63	-6.3	39.69
69	-0.3	0.09
73	3.7	13.69
71	1.7	2.89
$\sum x$ =693		$\sum (x-\bar{x})^2$ =234.10

Sample mean $(\bar{x}) = \frac{\sum x}{n} = \frac{693}{10} = 69.3$

Under the null hypothesis $H_o : \sigma^2 = 32$, the test statistic is

$$\chi^2 = \frac{ns^2}{\sigma^2} = \frac{234.10}{32} = 7.3156$$

which follows χ^2 distribution with (10-1)=9 d.f.

The table value of χ^2 at 5% level of significance is 16.9.

Since the calculated value of χ^2 is less than the tabulated value of χ^2 for 9 d.f at 5% level of significance. Hence H_o may be accepted.

CONDITIONS FOR THE VALIDITY OF CHI-SQUARE TEST

The chi-square test can be used if the following conditions are satisfied.

i. N i.e the number of observations must be sufficiently large otherwise the differences between the actual and observed frequencies would not be normally distributed.
ii. The sample observations should be independent.
iii. No theoretical cell frequency should be less than 5. If any theoretical cell frequency is less than 5, then for the application of χ^2 test, it is pooled with the preceeding or

succeeding frequency so that the pooled frequency is more than 5 and finally adjust for the d.f lost in pooling.

iv. The constraints on the cell frequencies, if any, should be linear (i.e they showed not involve square and higher powers of the frequencies) such as $\sum O_i = \sum E_i = N$.

2×2 Contingency table

For 2×2 table

a	b
c	d

Prove that chi-square test for independence gives

$$\chi^2 = \frac{N(ad-bc)^2}{(a+c)(b+d)(a+b)(c+d)}, \quad N = a+b+c+d$$

Solution: Under the hypothesis of independence of attributes

a	b	a+b
c	d	c+d
a+c	b+d	N

$$E(a) = \frac{(a+b)(a+c)}{N} \qquad (1)$$

$$E(b) = \frac{(a+b)(b+d)}{N} \qquad (2)$$

$$E(c) = \frac{(a+c)(c+d)}{N} \qquad (3)$$

$$E(d) = \frac{(b+d)(c+d)}{N} \qquad (4)$$

$$\therefore \ \chi^2 = \frac{[a-E(a)]^2}{E(a)} + \frac{[b-E(b)]^2}{E(b)} + \frac{[c-E(c)]^2}{E(c)} + \frac{[d-E(d)]^2}{E(d)} \qquad (A)$$

Now $(a-E(a)) = a - \frac{(a+b)(a+c)}{N}$ using (1)

$$= \frac{Na-(a^2+ac+ab+bc)}{N}$$

$$= \frac{a(a+b+c+d)-\left(a^2+ac+ab+bc\right)}{N} \qquad (\because\ N = a+b+c+d)$$

$$= \frac{a^2+ab+ac+da-a^2-ac-ab-bc}{N}$$

$$(a-E(a)) = \frac{ad-bc}{N}$$

Similarly, we will get

$$(b-E(b)) = \frac{ad-bc}{N},\ (c-E(c)) = \frac{ad-bc}{N} \text{ and } (d-E(d)) = \frac{ad-bc}{N}$$

Substituting these values in (A), we get

$$\chi^2 = \frac{(ad-bc)^2}{N^2}\left[\frac{1}{E(a)}+\frac{1}{E(b)}+\frac{1}{E(c)}+\frac{1}{E(d)}\right]$$

$$= \frac{(ad-bc)^2}{N^2}\left[\frac{N}{(a+b)(a+c)}+\frac{N}{(a+b)(b+d)}+\frac{N}{(a+c)(c+d)}+\frac{N}{(b+d)(c+d)}\right]$$

$$= \frac{(ad-bc)^2}{N^2}\left[\left(\frac{1}{(a+b)(a+c)}+\frac{1}{(a+b)(b+d)}\right)+\left(\frac{1}{(a+c)(c+d)}+\frac{1}{(b+d)(c+d)}\right)\right]$$

$$= \frac{(ad-bc)^2}{N^2}\left[\left(\frac{a+b+c+d}{(a+b)(a+c)(a+b)(b+d)}+\frac{b+d+a+c}{(a+c)(c+d)(b+d)}\right)\right]$$

$$\chi^2 = \frac{(ad-bc)^2}{N^2}\left[\frac{N}{(a+b)(a+c)(a+b)(b+d)}+\frac{N}{(a+c)(c+d)(b+d)}\right] \qquad (\because a+b+c+d=N)$$

$$= (ad-bc)^2\left[\frac{a+b+c+d}{(a+b)(a+c)(a+b)(b+d)}\right]$$

$$\chi^2 = \frac{N(ad-bc)^2}{(a+b)(a+c)(a+b)(b+d)}$$