UNIT V – CHI-SQUARE TEST

The $\chi 2$ 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 ~ N (μ, σ^2), then z = $z = \frac{X-\mu}{\sigma}$ ~ N(0, 1) and $z^2 = (\frac{X-\mu}{\sigma})^2$, is a chi-square variate with 1 degree of freedom abbreviated by letter $\chi 2$ of the Greek alphabet.

In general, if X_1, X_2, ---, X_n are n independent normal variate with means μ_1, μ_2, ---, μ_n and standard deviation $\sigma_1, \sigma_2 \ldots \sigma_n$ respectively then the variate

$$\chi 2 = (\frac{X_1 - \mu_1}{\sigma_1})^2 + (\frac{X_2 - \mu_2}{\sigma_2})^2 + \ldots\ldots + (\frac{X_n - \mu_n}{\sigma_n})^2 \quad = \sum(\frac{X_i - \mu_i}{\sigma_i})^2.$$

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

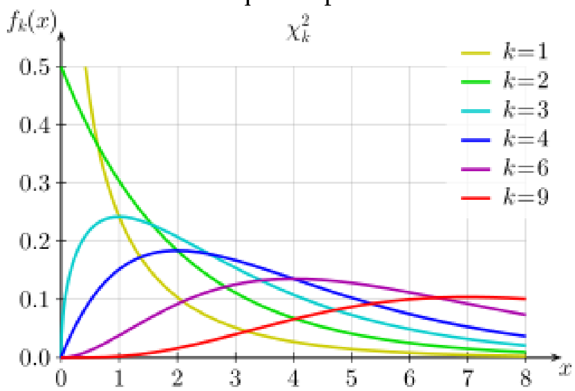
CHI-SQUARE DISTRIBUTION

In probability theory and statistics, the chi-squared distribution (also chi-square or χ2-distribution) with k degrees of freedom is the distribution of a sum of the squares of k independent standard normal random variables. The chi-squared distribution is a special case of the gamma distribution and is one of the most widely used probability distributions in inferential statistics, notably in hypothesis testing or in construction of confidence intervals. When it is being distinguished from the more general non central chi-squared distribution, this distribution is sometimes called the central chi-squared distribution.

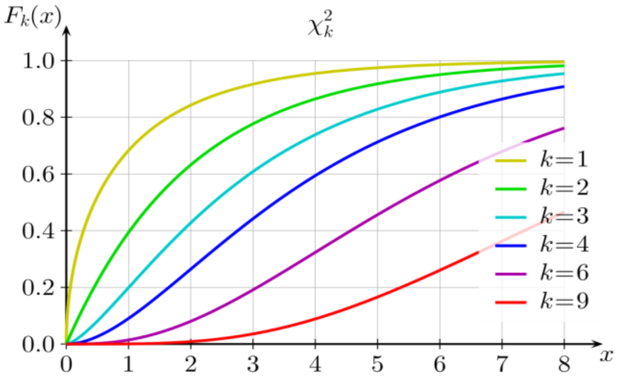
The chi-squared distribution is used in the common chi-squared tests for **goodness of fit** of an observed distribution to a theoretical one, the **independence of two criteria** of classification of qualitative data, and in confidence interval estimation for a population standard deviation of a normal distribution from a sample standard deviation. Many other statistical tests also use this distribution, such as Friedman's analysis of variance by ranks.

The χ2-distribution is not symmetrical and all the values are positive. For making use of this distribution, one is required to know the degrees of freedom since for different degrees of freedom we have different curves. The smaller the number of degrees of freedom, the more skewed is the distribution which is illustrated in Fig.

Chi-squared pdf



Chi-squared cdf



The Sampling distribution of chi-square can be closely approximated by a continuous normal curve as long as the sample size remains large. The probability function of Chi-square can be given as:

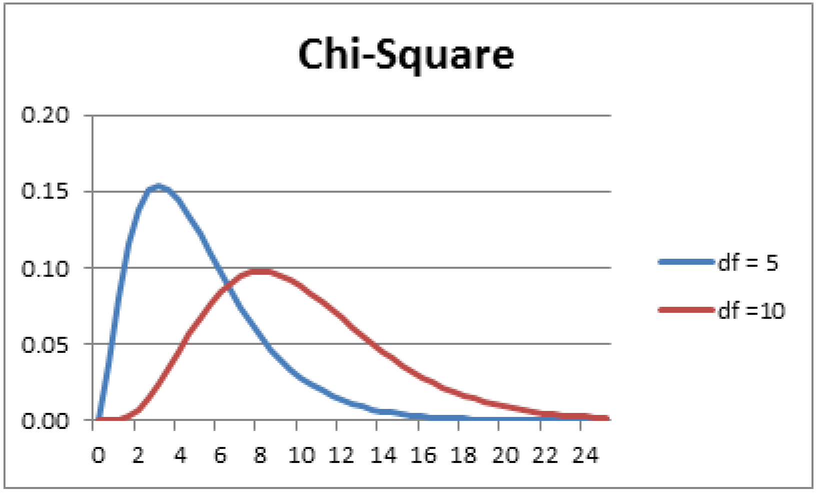
The **chi-square distribution** with *k* **degrees of freedom**, abbreviated $\chi^2(k)$, has probability density function

$$f(x) = \frac{x^{\frac{k}{2}-1} e^{-\frac{x}{2}}}{2^{k/2}\Gamma\left(\frac{k}{2}\right)}$$

k does not have to be an integer and can be any positive real number.
The chi-square distribution is the **gamma distribution** where *α* = *k*/2 and *β*= 2.
The **gamma distribution** has probability density function (pdf) given by

$$f(x) = \frac{x^{\alpha-1} e^{-x/\beta}}{\beta^{\alpha}\Gamma(\alpha)}$$

Through this, it is clear that the chi-square has only one parameter, i.e. degrees of freedom.
The following are the graphs of the pdf with degrees of freedom *df* = 5 and 10. As *df* grows larger the fat part of the curve shifts to the right and becomes more like the graph of a normal distribution.



Properties of Chi-Square Distribution

1. The chi-square distribution is a continuous probability distribution with the values ranging from **0 to ∞ (infinity)** in the positive direction. The **χ2** can never assume negative values.
2. The **shape of the chi-square distribution** depends on the number of **degrees of freedom** 'k'. When 'k' is small, the shape of the curve tends to be skewed to the right, and as the 'k' gets larger, the shape becomes more symmetrical and can be approximated by the normal distribution.
3. The mean of the chi-square distribution is equal to the degrees of freedom, i.e. **E(χ^2) = 'k'.** While the variance is twice the degrees of freedom, Viz. **n(χ^2) = 2k**. The mean of the distribution is equal to the number of degrees of freedom: μ = *k*. The variance is equal to two times the number of degrees of freedom: $\sigma^2 = 2 * k$
4. Median = $k(1-\frac{2}{9k})^3$ for large *df*
5. Mode = max (*df* – 1, 0) for *df* > 2
6. Range = [0.∞) for df >1 and Range = (0, ∞) for df=1
7. Skewness = $\sqrt{8/k}$
8. Kurtosis = 12/*k*
9. The χ2 distribution approaches the normal distribution as k gets larger with mean k and standard deviation as √2k. It has been determined that quantity **√2k** gives a better

approximation to normality than the χ^2 itself if the values are about 30 or more. Thus, the mean and standard deviation of the distribution of **√2χ^2 is equal to √2k-1 and one respectively.**

10. The **sum of independent χ^2 is itself a χ^2 variate**. Suppose, χ_1^2 is a χ^2 variate with degrees of freedom k_1 and χ_2^2 is another χ^2 variate with degrees of freedom k_2, then their sum $\chi_1^2 + \chi_2^2$ will be equal to χ^2 variate with $k_1 + k_2$ degrees of freedom. This property is called as the **additive property of Chi-square**.

Thus, χ^2 distribution depends on the degrees of distribution as its shape changes with the change in the 'k', and as 'k' becomes greater, χ^2 gets approximated by the normal distribution.

CHI-SQUARE TEST (χ^2 - test)

Chi-square tests: (**the test of goodness of fit, the test of independence, and the test of homogeneity**) are used to analyze categorical responses.

For all three tests the data are generally presented in the form of a contingency table (a rectangular array of numbers in cells). All three tests are based on the Chi-Square statistic:

The value of chi-square describes the magnitude of difference between observed frequencies and expected frequencies under certain assumptions. χ^2 value (χ^2 quantity) ranges from zero to infinity. It is zero when the expected frequencies and observed frequencies completely coincide. So greater the value of χ^2, greater is the discrepancy between observed and expected frequencies.

χ^2 -test is a statistical test which tests the significance of difference between observed frequencies and corresponding theoretical frequencies of a distribution **without any assumption about the distribution of the population.** This is one of the simplest and most widely used nonparametric test in statistical work. This test was developed by Prof. Karl Pearson in 1990.

Uses of χ^2 - test

The uses of chi-square test are:-

1. Useful for the test of goodness of fit:- χ^2 - test can be used to test whether there is goodness of fit between the observed frequencies and expected frequencies.

2. Useful for the test of independence of attributes:- χ^2 test can be used to test whether two attributes are associated or not.

3. Useful for the test of homogeneity:- χ^2 -test is very useful to test whether two attributes are homogeneous or not.

4. Useful for testing given population variance:- χ^2 -test can be used for testing whether the given population variance is acceptable on the basis of samples drawn from that population.

χ^2 -test as a test of goodness of fit:

As a **non-parametric test**, χ^2 -test is mainly used to test the goodness of fit between the observed frequencies and expected frequencies.

Procedure:-

1. Set up null hypothesis that there is goodness of fit between observed and expected frequencies.
2. Find the χ^2 value using the following formula:-

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

Where O = Observed frequencies

E = Expected frequencies

3. Compute the degree of freedom.

d. f. = n – r – 1

Where 'r' is the number of independent constraints to be satisfied by the frequencies

4. Obtain the table value corresponding to the level of significance and degrees of freedom.

5. Decide whether to accept or reject the null hypothesis. If the calculated value is less than the table value, we accept the null hypothesis and conclude that there is goodness of fit. If the calculated value is more than the table value we reject the null hypothesis and conclude that there is no goodness of fit.

The **goodness of fit test** answers the question, "Do the data fit well compared to a specified distribution?" This test considers one categorical variable and assesses whether the proportion of sampled observations falling into each category matches well enough to the null distribution for the given problem. For instance, the null distribution might be specified by a manufacturer, a product label, or the results of a previous study. The null hypothesis for the goodness of fit test specifies this null distribution which describes the population proportion of observations in each category.

The **test of homogeneity** answers the question, "Do two or more populations have the same distribution for one categorical variable?" This test considers one categorical variable and assesses whether this variable is distributed the same in two (or more) different populations. The null hypothesis for the test of homogeneity is that the distribution of the categorical variable is the same for the two (or more) populations.

The **test of independence** answers the question, "Are two factors (or variables) independent for a population under study?" This test considers two categorical variables and assesses whether there is a relationship between these two variables for a single population. The null hypothesis for the test of independence is that the two categorical variables are independent (that is, they are not related) for the population of interest.

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_ι (ι = 1, 2… κ) is a set of observed frequencies and E_ι ($\iota = 1,2,3,\ldots\kappa$) is the corresponding set of expected (theoretical) frequencies, then the Statistic χ 2 may be defined as

$\chi 2 = \sum[\frac{O_{i-} E_i}{E_i}]^2$, ($\sum O_i = \sum E_i$) 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.

Step1: H$_0$: there is no significant difference between observed frequency and expected frequency (it is hypothesised that die is unbiased)

H$_1$: there is significant difference between observed frequency and expected frequency (it is hypothesised that die is biased)

Step 2: α = 0.05 (Let us assume that significance level is 5%)

Step 3: 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

Observed frequecy (O)	Expected frequency (E)	(O – E)^2	$\frac{(O-E)^2}{E}$
16	22	36	1.64
20	22	4	0.18
25	22	9	.41
14	22	64	2.91
29	22	49	2.23
28	22	36	1.64
			$\sum\frac{(O-E)^2}{E} = 9.01$

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

Step 6: 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 falls in acceptance zone, so null hypothesis is accepted and hence we conclude there is no evidence against the hypothesis that die is unbiased.

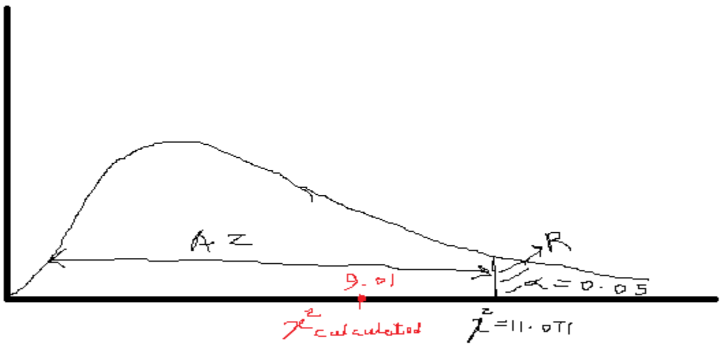


Illustration: The theory predicts the proportion of beans in the four groups A, B, C and D should be 9:3:3:1. In an experiment among 1600 beans, the numbers in the four groups were 882, 313, 287 and 118. Does the experimental result support theory?

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

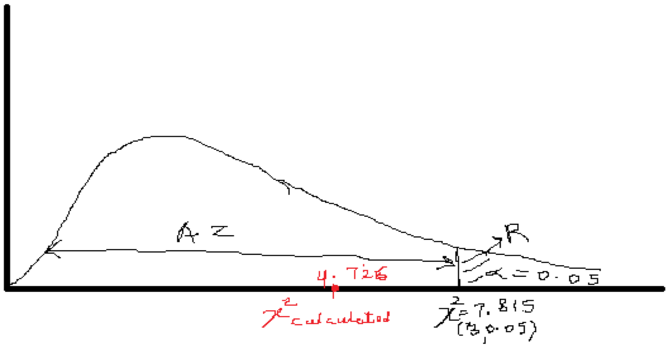
E(882) = $\frac{9}{16} \times 16000 = 900$, E(313) = $\frac{3}{16} \times 16000 = 300$, E(287) = $\frac{3}{16} \times 16000 = 300$, E(118) = $\frac{1}{16} \times 16000 = 100$,

Apply χ^2 test

O	E	$(O - E)^2$	$\frac{(O-E)^2}{E}$
882	900	324	.36
313	300	169	.5633
287	300	169	.5633
118	100	324	3.24
			$\sum \frac{(O-E)^2}{E} = 4.7226$

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.


A. Z
R
$\alpha = 0.05$
$\chi^2_{calculated}$
$\chi^2 = 7.815$
(3, 0.05)