PROBABILITY & STATISTICS

UNITS – III & IV: Testing of Hypothesis-I & Testing of Hypothesis-II

Estimation:

Def: The 'Population' in a statistical study is the set or collection or totality of observations about which inferences are to be drawn.

Ex: (i) Engineering graduate students in India.

(ii) Total Production of items in a month from a factory.

(iii) Budget of India

Def: 'Sample' is a finite subset of population.

Ex: (i) Engineering graduate students in India (Population). Engineering graduate students in Karnataka (Sample)

(ii) Total Production of items in a month from a factory (Population). Total production of items in a day (Sample).

(iii) Budget of India (Population). Budget of a state (Sample).

Def: A population parameter is a statistical measure or constant obtained from the population.

Ex: Population mean (μ), Population variance (σ^2).

Def: A sample statistic is a statistical measure computed from sample observations.

Ex: Sample mean ($\bar{x}$), Sample variance (S^2)

Def: The process of drawing or obtaining samples is called sampling. If $n \geq 30$, the sampling is known as large sampling and if $n < 30$, the sampling is known as small sampling or exact sampling.

Sampling Distribution:

Consider a finite population of size N. Draw all possible samples of size n from this population. The total number of samples drawn is given by $NC_n = \frac{N!}{n!(N-n)!} = k$ (say). Compute a statistic 'S' (such as mean, standard deviation, median, mode etc.) for each of these samples. Let these statistic values be $\{s_1, s_2, ...s_k\}$. Then the sampling distribution of the statistic 'S' is the set of values $\{s_1, s_2, ...s_k\}$ of the statistic 'S' together with the samples.

i.e., the sampling distribution of 'S' is

Sample number	1	2	3	…	k
Statistic	s_1	s_2	s_3	…	s_k

Note: Thus sampling distribution describes how a statistic 'S' will vary from one sample to the other sample of same size.

Note: If the statistic 'S' is mean, then the corresponding distribution is known as sampling distribution of mean. Thus, if 'S' is variance, proportion or median etc., the associated distribution is known as sampling distribution of variance, sampling distribution of proportion etc.

Standard error:

The standard deviation of the sampling distribution of a statistic is known as standard error of that statistic.

S. No.	Sampling distribution	Standard error	Statistic
1	Means ($\bar{x}$)	$\frac{\sigma}{\sqrt{n}}$	$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$
2	Proportions (P)	$\sqrt{\frac{PQ}{n}}$	$z = \frac{p - P}{\sqrt{PQ/n}}$
3	Difference of means ($\bar{x}_1 - \bar{x}_2$)	$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$	$z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$
4	Difference of Proportions(P_1-P_2)	$\sqrt{\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}}$	$z = \frac{P_1 - P_2}{\sqrt{\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}}}$

Note: The standard normal varaite of any statistic 'S'will be, $Z = \frac{s - E(s)}{s.e(S)}$, E(S) is expectation of S and s.e(S) is Standard error of 'S'.

Probable error:

The probable error for any sampling distribution of a statistic is 0.6745 times of the standard error.

Statistical Estimation: Statistical estimation is a part of statistical inference, where a population parameter is estimated from the corresponding sample statistic.

Estimation may be divided into two types. (i) Point Estimation (ii) Interval Estimation

Point Estimation: Point estimation is an estimation in which the population parameter will be estimated by a single number.

For example, if we use a value $\bar{x}$ to estimate the mean of population (or) a value s^2 to estimate a population variance, we are in each case using a point estimate of the parameter in question. These estimates are called point estimates.

Interval Estimation: Interval estimation is an estimation in which the population parameter will be estimated by two numbers between which the parameter is considered to lie.

Ex: For a trip from Bengaluru to Hyderabad by car, we estimate the distance as 600 km, mileage/litre as 12 km, price/litre of fuel as Rs. 70/- from which eventually estimate the entire cost of the trip. Also we might estimate the distance being between 550 to 650 km, mileage/litre of fuel as 11- 14 km, price/litre as Rs. 65/- - 75/-. In the first case, we are estimating distance, mileage, cost of fuel as specific values or points. So this method of estimation is known as point estimation. Where as in the second case, the above parameters are estimated by intervals. So this method is known as interval estimation.

Unbiased Estimator: A statistic $\bar{\theta}$ is said to be an unbiased estimator, or its value an unbiased estimate, of the population parameter, if and only if the mean or expectation of the sampling distribution of the statistic is equal to the population parameter θ.

i.e., $\bar{\theta}$ is an unbiased estimator of θ if $E(\bar{\theta}) = \theta$.

More Efficient Unbiased Estimator:

Let $\bar{\theta_1}$, $\bar{\theta_2}$ be two unbiased estimators of θ and let σ_1^2, σ_2^2 are variances of their sampling distributions. $\bar{\theta_1}$ is said to be more efficient estimator of θ if $\sigma_1^2 < \sigma_2^2$.

Maximum Error of Estimate:

When we use a sample statistic to estimate the population parameter, the chances are less that the estimate will actually equal the parameter. Hence, it would seem desirable to accompany such point estimate with some statement as to how close we might reasonably expect the estimate to be.

The error is the difference between the estimator and quantity it is supposed to estimate.

In the case population mean:

We know that for large n, $Z = \dfrac{\bar{x} - \mu}{\sigma / \sqrt{n}}$ is a random variable having approximately the standard normal distribution. We can assert with probability $1-\alpha$ the inequality $-Z_{\alpha/2} < \dfrac{\bar{x} - \mu}{\sigma / \sqrt{n}} < Z_{\alpha/2}$, where $Z_{\alpha/2}$ is such that the normal curve area to its right equals to $\alpha/2$.

Let E be the Maximum error of $|\bar{x} - \mu|$, i.e., maximum error of estimator μ, given by E = $-Z_{\frac{\alpha}{2}} \dfrac{\sigma}{\sqrt{n}}$ with probability $1-\alpha$.

Confidence Interval for the Mean:

Suppose that a population has mean μ and variance σ^2. A random sample size $n \geq 30$ is taken from this population. The sample mean $\bar{x}$ is a reasonable point estimate of

the unknown mean μ. A $100(1-\alpha)$ % confidence interval on μ can be obtained by considering the sampling distribution of the sample mean $\bar{x}$.

The standard normal variate of the statistic is, $Z = \dfrac{\bar{x} - \mu}{\sigma / \sqrt{n}}$.

Let $P\left(-Z_{\alpha/2} < \dfrac{\bar{x} - \mu}{\sigma / \sqrt{n}} < Z_{\alpha/2} \right) = 1 - \alpha$

The confidence interval for the mean is given by $\left(\bar{x} - Z_{\frac{\alpha}{2}} \dfrac{\sigma}{\sqrt{n}},\ \bar{x} + Z_{\frac{\alpha}{2}} \dfrac{\sigma}{\sqrt{n}} \right)$

TESTING OF HYPOTHESIS

Tests of Significance: A very important aspect of the sampling theory is the study of the *tests of significance,* which enable us to decide on the basis of the sample results, if

(i) the deviation between the observed sample statistic and the hypothetical parameter value, or

(ii) the deviation between two independent sample statistics;

is significant or might be attributed to chance or the fluctuations of sampling.

Def: A statistical hypothesis is a statement about the parameters of one or more populations.

Ex: (i) The majority of the men in a city are smokers.

(ii) The teaching methods in both the institutions are effective.

There are two types of hypothesis: 1) Null Hypothesis 2) Alternative Hypothesis.

Null Hypothesis:

For applying the tests of significance, we first setup a hypothesis – a definite statement about the population parameter. Such a hypothesis, which is usually a hypothesis of no difference, is called Null hypothesis and is usually denoted by H_0.

For example, in case of a single statistic, H_0 will be that the sample statistic does not differ significantly from the hypothetical parameter value and in the case of two statistics; H_0 will be that the sample statistics do not differ significantly.

Having set up the null hypothesis we compute the probability P that the deviation between the observed sample statistic and the hypothetical parameter value might have occurred due to fluctuations of sampling. If the deviation comes out to be significant (as measured by a test of significance), null hypothesis is refuted or rejected at the particular level of significance adopted and if the deviation is not significant, null hypothesis may be retained or accepted at that level.

Alternative Hypothesis:

Any hypothesis which is complementary to the null hypothesis is called an alternative hypothesis, usually denoted by H_1. For example, if we want to test the null

hypothesis that the population has a specified mean μ_0, (say), *i.e., H0:* $\mu = \mu_0$, then the alternative hypothesis could be

(i) H_1: $\mu \neq \mu_0$ (ii) H_1: $\mu > \mu_0$ (iii) H_1: $\mu < \mu_0$

The alternative hypothesis in (i) is known as a two *tailed alternative* and the alternatives in (ii) and *(iii)* are known as *right tailed* and *left-tailed alternatives* respectively.

The setting of alternative hypothesis is very important since it enables us to decide whether we have to use a single-tailed (right or left) or two tailed test.

Errors in Sampling:

The main objective in sampling theory is to draw valid inferences about the population parameters on the basis of the sample results. In practice we decide to accept or to reject the population after examining a sample from it. As such we have two types of errors.

(i) Type I error: Reject H_0 when it is true.

(ii) Type II error: Accept H_0 when it is wrong

Critical Region:

A region corresponding to a statistic 't' in the sample space 'S' which leads to rejection of H_0 is called Critical region or rejection region. The region which leads to the acceptance of H_0 is called Acceptance region.

If we write, P(Reject H_0 when it is true) = P(Type I error) = α

P(Accept H_0 when it is wrong) = P(Type II error) = β

then α and β are called sizes of Type I and Type II errors respectively.

Power of the test:

The probability of type II error is denoted by β and 1-β is called the power of the test of the hypothesis.

Level of Significance:

The probability of type I error is called the level of significance. The levels of significance usually employed in testing of hypothesis are *5%* and 1 %. The level of significance is always fixed in advance before collecting the sample information.

Table: Critical values of Z

Level of Significance α	1% (α=0.01)	5% (α=0.05)	10% (α=0.1)
Two tailed test	$\lvert Z_\alpha \rvert$=2.58	$\lvert Z_\alpha \rvert$=1.96	$\lvert Z_\alpha \rvert$=1.645
Right tailed test	Z_α=2.33	Z_α=1.645	Z_α=1.28
Left tailed test	Z_α=-2.33	Z_α=-1.645	Z_α=-1.28

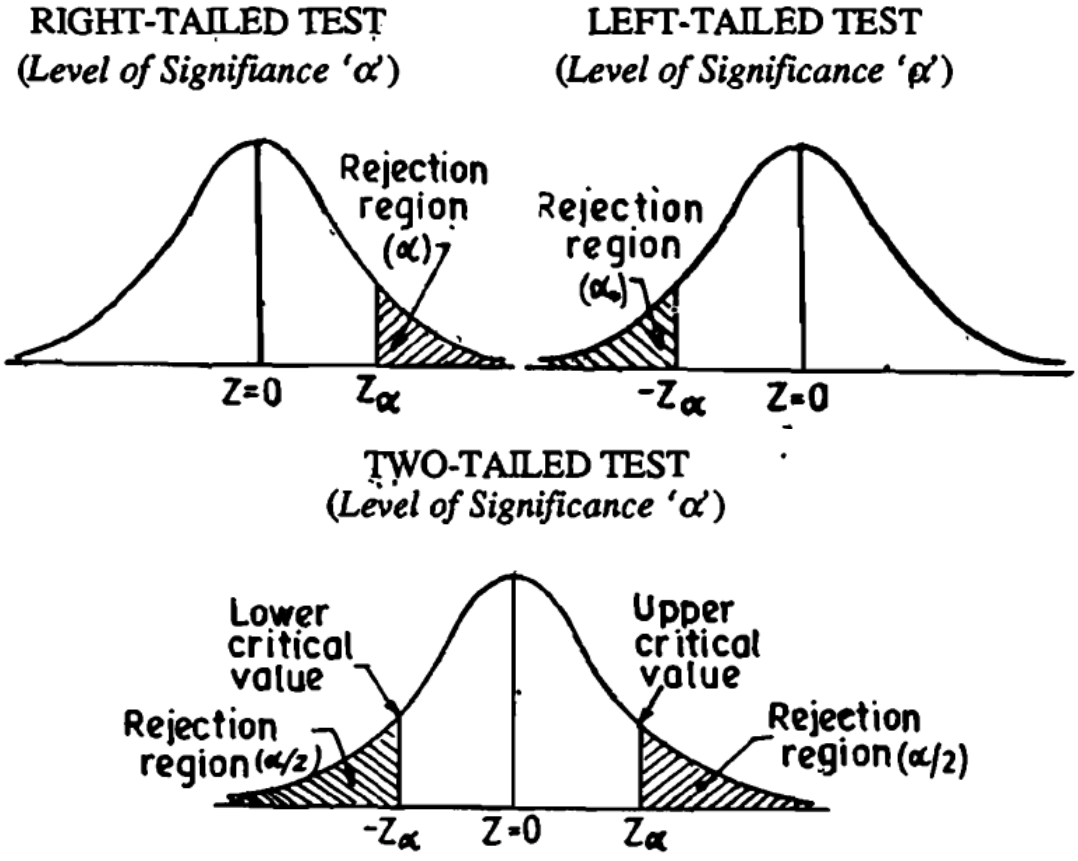
One – Tailed and Two – Tailed Tests:

In any statistical test, the critical region is represented by a portion of the area under the probability curve of the sampling distribution of the test statistic.

A test of any statistical hypothesis where the alternative hypothesis is one tailed (right tailed or left tailed) is called a *one tailed test.* For example, a test for testing the mean of a population $H_0 : \mu = \mu_0$ against the alternative hypothesis, $H_1 : \mu > \mu_0$ (Right tailed) or $H_1 : \mu < \mu_0$ (Left tailed) is a single tailed test. In the right tailed rest $(H_1 : \mu > \mu_0)$, the critical region lies entirely in the right tail of the sampling distribution of $\bar{x}$, while for the left tail test $(H_1 : \mu < \mu_0)$, the critical region is entirely in the- left tail of the distribution.

A test of statistical hypothesis where the alternative hypothesis is two tailed such as: $H_0 : \mu = \mu_0$, against the alternative hypothesis $H_1 : \mu \neq \mu_0$, ($\mu > \mu_0$ and $\mu < \mu_0$), is known as *two tailed test* and in such a case the critical region is given by the portion of the area lying in both the tails of the probability curve of the test statistic.

In a particular problem, whether one tailed or two tailed test is to be applied depends entirely on the nature of the alternative hypothesis. If the alternative hypothesis is two-tailed we apply two-tailed test and if alternative hypothesis is one-tailed, we apply one tailed test.



Procedure for Testing of Hypothesis:

Various steps involved in testing of Hypothesis are given below

Step 1: Null Hypothesis: Define or set up a null Hypothesis H_0 taking into consideration the nature of the problem and data involved.

Step 2: Alternative Hypothesis: Set up the Alternative Hypothesis H_1 so that we could decide whether we should use one-tailed or two-tailed test.

Step 3: Level of Significance: Select the appropriate level of significance (α) depending on the reliability of the estimates and permissible risk. That is a suitable α is selected in advance if it is not given in the problem. (Usually we chose 5% level of significance)

Step 4: Test Statistic: Compute the test statistic $Z = \dfrac{t - E(t)}{S.E\ of\ t}$ under the null hypothesis.

Here t is a sample statistic and S.E is the standard error of t.

Step 5: Conclusion: We compare the computed value of the test statistic Z with the critical value Z_α at given level of significance(α).

If $|Z| < Z_\alpha$, (that is, if the absolute value of the calculated value of Z is less than the critical value Z_α) we conclude that it is not significant. We accept the null hypothesis.

If $|Z| > Z_\alpha$ then the difference is significant and hence the null hypothesis is rejected at the level of significance α.

SMALL SAMPLE TESTS

Introduction:

In the earlier chapter, we considered certain tests of significance based on the theory of the normal distribution. The assumptions made in deriving those tests will be valid only for large samples. In case samples are small (n<30), we can formulate tests of hypotheses and significance using other distributions besides the normal, such as Student's t, Chi-square, F, etc.

Degrees of freedom (d.f.)

The number of independent variables which make up the statistic is known as the degree of freedom and it is denoted by ν (the letter "Nu" of the Greek alphabet). In other words, it is the number of values in a set of data which may be assigned arbitrarily or, it refers to the number of "independent constraints" in a set of data.

Note: Degree of freedom is a number which indicates how many of the values of a variable may be independently (or freely) chosen.

Example: If we have to choose any four numbers freely, then we may choose 12, 7, 19, 85 or any other set of four numbers. In this case, all the four numbers have freedom to vary; we say that the degree of freedom is 4. If we impose a restriction on the numbers, say the

sum is 50, then we can choose first 3 numbers freely and the fourth number is such that the sum is 50. Thus the 3 variables are free and independent choices for finding the fourth. Hence, the degree of freedom is 3.

***t* - Distribution (or) Student's *t* - Distribution:**

It is used for testing of hypothesis when the sample size is small and population S.D. σ is not known.

Definition: If $\{x_1, x_2, ..., x_n\}$ be any random sample of size n drawn from a normal population with mean μ and variance σ^2, then the test statistic is defined by $t = \frac{\bar{x} - \mu}{S/\sqrt{n}}$

where $\bar{x}$=sample mean and $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$ is an unbiased estimate of σ^2. The test statistic $t = \frac{\bar{x} - \mu}{S/\sqrt{n}}$ is a random variable having the t- distribution with $\nu = n-1$ degrees of freedom and with probability density function $f(t) = y_0\left(1 + \frac{t^2}{\nu}\right)^{\frac{\nu+1}{2}}$ where $\nu = n-1$ and y_0 is a constant got by $\int_{-\infty}^{\infty} f(t)dt = 1$. This is known as Student's t- Distribution or simply t- distribution.

Properties of t - Distribution:

1. The shape of t- distribution is bell-shaped, which is similar to that of a normal distribution and is symmetrical about the mean.
2. The t- Distribution curve is also asymptotic to the t-axis, i.e., the two tails of the curve on both sides of $t = 0$ extends to infinity
3. It is symmetrical about the line $t = 0$
4. The form of the probability curve varies with degrees of freedom i.e., with sample size.
5. It is unimodal with Mean=Median=Mode.
6. The mean of standard normal distribution and as well as t- distribution is zero but the variance of t- distribution depends upon the parameter ν which is called the degrees of freedom.
7. The variance of t- distribution exceeds 1, but approaches 1 as $n \to \infty$. Infact the t- distribution with ν-degrees of freedom approaches standard normal distribution as $\nu = (n-1) \to \infty$.

Chi-Square (χ^2) Distribution

Chi-squared distribution is a continuous probability distribution of a continuous random variable X with probability density function given by

$$f(x)=\begin{cases}\dfrac{1}{2^{\frac{\nu}{2}}\Gamma\left(\frac{\nu}{2}\right)}x^{\frac{\nu}{2}-1}e^{-\frac{x}{2}}, & for \quad x>0\\ 0, & otherwise\end{cases}$$

where ν is a positive integer which is the only single parameter of the distribution, also known as "degrees of freedom".

χ^2-distribution was extensively used as a measure of goodness of fit and to test the independence of attributes.

Properties of χ^2-distribution:

1. χ^2- distribution curve is not symmetrical, lies entirely in the first quadrant and hence not a normal curve, since χ^2 varies from 0 to ∞.
2. It depends only on the degrees of freedom ν.
3. If χ_1^2 and χ_2^2 are two-independent distributions with ν_1 and ν_2 degrees of freedom, the $\chi_1^2+\chi_2^2$ will be Chi-square distribution with $\nu_1+\nu_2$ degrees of freedom. That is, it is additive.
4. χ_α^2 represents the χ^2 value such that the area under the chi-square curve to its $\left(\chi_\alpha^2\right)$ right is equal to α.
5. For a Chi-square distribution, Mean= ν and variance = 2ν

F- Distribution: (Sampling Distribution of the Ratio of the Sample Variances)

Another important continuous probability distribution which plays an important role in connection with sampling from normal populations is the F-distribution. This is used to determine whether the two samples come from two populations having equal variances.

Let S_1^2 be the sample variance of an independent sample of size n_1 drawn from a normal population of variance σ_1^2. Similarly, S_2^2 be the sample variance in an independent sample of size n_2 drawn from other population of variance σ_2^2.

Consider the sampling distribution of the ratio of the variances of the two independent random samples defined by $F=\dfrac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}=\dfrac{\sigma_2^2S_1^2}{\sigma_1^2S_2^2}$, which follows F-distribution with $\nu_1=n_1-1$ and $\nu_2=n_2-1$ degrees of freedom.

F-distribution can be used to test the equality of several population means, comparing sample variances and analysis of variance. F determines whether the ratio of two sample variances S_1^2 and S_2^2 is too small or too large. When F is close to 1, the two sample variances are almost same. In practice, it is customary, to take the larger sample variance as the numerator. F is always a positive number.

The sampling distribution of F is of the form $f(F) = K\frac{F^{(\nu_1-\nu_2)/2}}{(\nu_1 F+\nu_2)^{(\nu_1+\nu_2)/2}}$ where K is determined by $\int_0^\infty f(F)dF = 1$.

Properties of F-distribution:

(i) F- distribution is free from population parameters and depends upon degrees of freedom only.

(ii) F- distribution curve lies entirely in first quadrant.

(iii) The F-curve depends not only on the two parameters ν_1 and ν_2 but also on the order in which they are stated.

(iv) $F_{1-\alpha}(\nu_1,\nu_2) = \frac{1}{F_\alpha(\nu_2,\nu_1)}$, where $F_\alpha(\nu_1,\nu_2)$ is the value of F with ν_1 and ν_2 degrees of freedom such that the area under the F- distribution curve to the right of F_α is α.

(v) The mode of F – distribution is less than unity.

Student's t-test for single mean:

Suppose we want to test

(i) If a random sample of size n has been drawn from a normal population with a specified mean μ.

(ii) If the sample mean differs significantly from the hypothetical value μ of the population mean.

Let a random sample of size n (n<30) has a sample mean $\bar{x}$. To test the hypothesis that the population mean μ has a specified value μ_0, when population standard deviation σ is not known.

Let the Null hypothesis be $H_0: \mu = \mu_0$, then the alternative hypothesis could be

(i) $H_1: \mu \neq \mu_0$ (ii) $H_1: \mu > \mu_0$ (iii) $H_1: \mu < \mu_0$

The alternative hypothesis in (i) is known as a two *tailed alternative* and the alternatives in (ii) and *(iii)* are known as *right tailed* and *left-tailed alternatives* respectively.

Assuming that H_0 is true, the test statistic given by $t = \frac{\bar{x} - \mu}{s/\sqrt{n-1}}$, where s is the sample standard deviation follows $t-$distribution with $\nu = (n-1)$ degrees of freedom.

We calculate the value of $|t|$ and compare this value with the table value of t at α level of significance. If the calculated value of $t >$ the table value of t, we reject H_0 at α level. Otherwise we accept H_0. For a two tailed test at α level of significance, value of $\frac{\alpha}{2}$ is taken for α

Student t-test for difference of means:

Let $\bar{x}$ and $\bar{y}$ be the means of two independent samples of sizes n_1 and n_2 $(n_1 < 30, n_2 < 30)$ drawn from two normal populations having means μ_1 and μ_2.

To test whether the two population means are equal (i.e., to test whether the difference is $\mu_1 - \mu_2$ significant),

Let the Null Hypothesis be $H_0 : \mu_1 = \mu_2$, then the Alternative Hypothesis is $H_1 : \mu_1 \neq \mu_2$. If $\sigma_1 = \sigma_2 = \sigma$, then an unbiased estimate S^2 of the common variance σ^2 is given by $S^2 = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}$ where s_1^2 and s_2^2 are the two sample variances.

Also Standard error of $\left(\bar{x} - \bar{y}\right) = S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$ where $S = \sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}}$

Assuming that H_0 is true, the test statistic t given by $t = \frac{\bar{x} - \bar{y}}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$, follows t - distribution with $(n_1 + n_2 - 2)$ degrees of freedom.

Here $\bar{x} = \frac{1}{n_1}\sum_{i=1}^{n_1} x_i$, $\bar{y} = \frac{1}{n_2}\sum_{i=1}^{n_2} y_i$ and $S^2 = \frac{1}{n_1 + n_2 - 2}\left[\sum\left(x_i - \bar{x}\right)^2 + \sum\left(y_i - \bar{y}\right)^2\right]$

Or $S^2 = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}$

We calculate the value of $|t|$ and compare this value with the table value of t at α level of significance. If the calculated value of $|t| >$ the table value of t, we reject H_0 at α level. Otherwise we accept H_0.

Snedecor's F-test: If s_1^2 and s_2^2 are the variances of two samples of sizes n_1 and n_2 respectively, then the population variances are given by $n_1 s_1^2 = (n_1 - 1)S_1^2$ and $n_2 s_2^2 = (n_2 - 1)S_2^2$.

The quantities $\nu_1 = n_1 - 1$ and $\nu_2 = n_2 - 1$ are called degrees of freedom of these estimates. We want to test if these S_1^2 and S_2^2 are significantly different or if the samples may be regarded as drawn from the same population or from two populations with same variance σ^2.

Test for Equality of Two population variances:

Let two independent random samples of sizes n_1 and n_2 be drawn from two populations. To test the hypothesis that the two population variances σ_1^2 and σ_2^2 are equal.

Let the Null Hypothesis be $H_0 : \sigma_1^2 = \sigma_2^2$, then the Alternate hypothesis is $H_1 : \sigma_1^2 \neq \sigma_2^2$.

The estimates of σ_1^2 and σ_2^2 are given by $S_1^2 = \frac{n_1 s_1^2}{n_1 - 1} = \frac{\sum (x_i - \bar{x})^2}{n_1 - 1}$ and $S_2^2 = \frac{n_2 s_2^2}{n_2 - 1} = \frac{\sum (y_i - \bar{y})^2}{n_2 - 1}$,

where s_1^2 and s_2^2 are the variances of two samples.

Assuming H_0 is true, the test statistic $F = \frac{S_1^2}{S_2^2}$ or $\frac{S_2^2}{S_1^2}$ according as $S_1^2 > S_2^2$ or $S_2^2 > S_1^2$ follows F-distribution with $(n_1 - 1,\ n_2 - 1)$ degrees of freedom.

If the calculated value of F > the tabulated value of F at α level, we reject the Null hypothesis and conclude that the variances σ_1^2 and σ_2^2 are not equal. Otherwise we accept the Null hypothesis and conclude that σ_1^2 and σ_2^2 are equal.

Chi – Square (χ^2) Test:

Def: If a set of events $A_1, A_2, \ldots, A_n$ are observed to occur with frequencies $O_1, O_2, \ldots, O_n$ respectively and according to probability rules $A_1, A_2, \ldots, A_n$ are expected to occur with frequencies $E_1, E_2, \ldots, E_n$ respectively. Here, O_i's are called observed frequencies and E_i's are called expected frequencies. If O_i's and E_i's are known, then χ^2 is defined as

$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$ with (n-1) degrees of freedom.

χ^2 Test as a goodness of fit:

We use this test to decide whether the discrepancy between theory and experiment is significant or not i.e., to test whether the difference between the theoretical and observed values can be attributed to chance or not.

Let the Null hypothesis H_0 be that there is no significant difference between the observed values and the corresponding expected values. Then the Alternative hypothesis H_1 is that the above difference is significant.

Let $O_1, O_2, \ldots, O_n$ be a set of observed frequencies and $E_1, E_2, \ldots, E_n$ be the corresponding set of expected frequencies. Then the test statistic χ^2 is given by

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

Assuming that H_0 is true, the test statistic χ^2 follows Chi – square distribution with (n-1) degrees of freedom, where $\sum_{i=1}^{n} O_i = \sum_{i=1}^{n} E_i$

If the calculated value of χ^2 **>** tabulated value of χ^2 at α level, the Null hypothesis is rejected. Otherwise, it is accepted.

Chi-Square test for independence of attributes:

Literally, an attribute means a quality or characteristic. Examples of attributes are drinking, smoking, blindness, honesty, beauty etc.

An attribute may be marked by its presence (position) or absence in a number of a given population. Let the observations be classified according to two attributes and frequencies O_i in the different categories be shown in a two-way table, called contingency table.

We have to test on the basis of cell frequencies whether the two attributes are independent or not. We take the Null-hypothesis H_0 that there is no association between the attributes i.e., we assume that the two attributes are independent. The expected frequencies (E_i) of any cell = $\frac{\text{Row Total} \times \text{Column Total}}{\text{Grand Total}}$.

The test statistic $\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$ approximately follows Chi-square distribution with degrees of freedom = (Number of rows - 1)×(Number of columns - 1).

If the calculated value χ^2 of is less than the table value at a specified level of significance, the hypothesis holds good i.e., the attributes are independent and do not bear any association. On the other hand, if the calculated value of χ^2 is greater than the table value at a specified level of significance, we say that the results of the experiment do not support the hypothesis, in other words, the attributes are associated each other.

Example: Let us consider two attributes A and B which are divided in to two classes. The various cell frequencies can be expressed in the following table known as 2x2 contingency table.

Categories → / Attributes ↓	Category 1	Category 2	Row Total ↓
A	a	b	a+b
B	c	d	c+d
Column Total→	a+c	b+d	N=a+b+c+d

The expected frequencies are given by,

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

The value of χ^2 is given by $\chi^2=\frac{N(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)}$ where N=a+b+c+d with degrees of freedom (2-1)(2-1)=1.