Module II: Data Analysis – I:

Hypothesis testing; Z-test, t-test, F-test, chi-square test.

Analysis of variance (One and Two way).

Non-parametric, Test – Sign Test, Run test, Krushall – Wallis test

PARAMETRIC TESTS

The important parametric tests are:

(1) z-test;

(2) t-test;

(3) $\chi 2$-test, and

(4) F-test.

All these tests are based on the assumption of normality i.e., the source of data is considered to be normally distributed.

1) **z-test:**

z-test is based on the normal probability distribution and is used for judging the significance of several statistical measures, particularly the mean. z-test is generally used for comparing the mean of a sample to some hypothesized mean for the population in case of large sample, or when population variance is known. z-test is also used for judging the significance of difference between means of two independent samples in case of large samples, or when population variance is known. z-test is also used for comparing the sample proportion to a theoretical value of population proportion or for judging the difference in proportions of two independent samples when n happens to be large. Besides, this test may be used for judging the significance of median, mode, coefficient of correlation and several other measures.

2) **t-test**:

t-test is based on t-distribution and is considered an appropriate test for judging the significance of a sample mean or for judging the significance of difference

between the means of two samples in case of small sample(s) when population variance is not known (in which case we use variance of the sample as an estimate of the population variance). In case two samples are related, we use paired t-test (or what is known as difference test) for judging the significance of the mean of difference between the two related samples. It can also be used for judging the significance of the coefficients of simple and partial correlations. The relevant test statistic, t, is calculated from the sample data and then compared with its probable value based on t-distribution (to be read from the table that gives probable values of t for different levels of significance for different degrees of freedom) at a specified level of significance for concerning degrees of freedom for accepting or rejecting the null hypothesis. It may be noted that t-test applies only in case of small sample(s) when population variance is unknown.

3) χ^2 –test or chi- square test:

χ^2 *-test* is based on chi-square distribution and as a parametric test is used for comparing a sample variance to a theoretical population variance.

4) F – Test:

F-test is based on *F*-distribution and is used to compare the variance of the two-independent samples. This test is also used in the context of analysis of variance (ANOVA) for judging the significance of more than two sample means at one and the same time. It is also used for judging the significance of multiple correlation coefficients. Test statistic, *F*, is calculated and compared with its probable value (to be seen in the *F*-ratio tables for different degrees of freedom for greater and smaller variances at specified level of significance) for accepting or rejecting the null hypothesis.

Hypothesis Testing Of Means

Mean of the population can be tested presuming different situations such as the population may be normal or other than normal, it may be finite or infinite, sample size may be large or small, variance of the population may be known or unknown and the alternative hypothesis may be two-sided or one sided. Our testing technique will differ in different situations. We may consider some of the important situations.

1. *Population normal, population infinite, sample size may be large or small but variance of the population is known, Ha may be one-sided or two-sided*:
In such a situation *z*-test is used for testing hypothesis of mean and the test statistic *z* is worked our as under:

$$z = \frac{\bar{X} - \mu_{H_0}}{\sigma_p / \sqrt{n}}$$

2. *Population normal, population finite, sample size may be large or small but variance of the population is known, Ha may be one-sided or two-sided*:
In such a situation *z*-test is used and the test statistic *z* is worked out as under (using finite population multiplier):

$$z = \frac{\bar{X} - \mu_{H_0}}{\left(\sigma_p / \sqrt{n}\right) \times \left[\sqrt{(N-n)/(N-1)}\right]}$$

3. *Population normal, population infinite, sample size small and variance of the population unknown, Ha may be one-sided or two-sided:*
In such a situation *t*-test is used and the test statistic *t* is worked out as under:

$$t = \frac{\bar{X} - \mu_{H_0}}{\sigma_s / \sqrt{n}} \text{ with d.f.} = (n-1)$$

and

$$\sigma_s = \sqrt{\frac{\Sigma \left(X_i - \bar{X}\right)^2}{(n-1)}}$$

4. *Population normal, population finite, sample size small and variance of the population unknown, and Ha may be one-sided or two-sided:*
In such a situation *t*-test is used and the test statistic 't' is worked out as under (using finite population multiplier):

$$t = \frac{\bar{X} - \mu_{H_0}}{\left(\sigma_s / \sqrt{n}\right) \times \sqrt{(N-n)/(N-1)}} \text{ with d.f.} = (n-1)$$

and $$\sigma_s = \sqrt{\frac{\sum\left(X_i - \bar{X}\right)^2}{(n-1)}}$$

5. *Population may not be normal but sample size is large, variance of the population may be known or unknown, and Ha may be one-sided or two-sided:*
In such a situation we use *z*-test and work out the test statistic *z* as under:

$$z = \frac{\bar{X} - \mu_{H_0}}{\sigma_p/\sqrt{n}}$$

(This applies in case of infinite population when variance of the population is formula.)
Or,

$$z = \frac{\bar{X} - \mu_{H_0}}{\left(\sigma_p/\sqrt{n}\right) \times \sqrt{(N-n)/(N-1)}}$$

(This applies in case of finite population when variance of the population is known but when variance is not known, we use σ_S in place of σ_p in this formula.)

Example 1:

A sample of 400 male students is found to have a mean height 67.47 inches. Can it be reasonably regarded as a sample from a large population with mean height 67.39 inches and standard deviation 1.30 inches? Test at 5% level of significance.

Solution: Taking the null hypothesis that the mean height of the population is equal to 67.39 inches, we can write:

$$H_0: \mu_{H0} = 67.39''$$
$$H_a: \mu_{H0} \neq 67.39''$$

and the given information as $\bar{X} = 67.47''$, $\sigma_P = 1.30''$, $\nu = 400$. Assuming the population to be normal, we can work out the test statistic *z* as under:

$$z = \frac{\bar{X} - \mu_{H_0}}{\sigma_p/\sqrt{n}} = \frac{67.47 - 67.39}{1.30/\sqrt{400}} = \frac{0.08}{0.065} = 1.231$$

.

As *Ha* is two-sided in the given question, we shall be applying a two-tailed test for determining the rejection regions at 5% level of significance which comes to as under, using normal curve area table:

$$R: |z| > 1.96$$

The observed value of *z* is 1.231 which is in the acceptance region since *R:* | *z* | > 1.96 and thus *H*0 is accepted. We may conclude that the given sample (with mean height = 67.47″) can be regarded to have been taken from a population with mean height 67.39″ and standard deviation 1.30″ at 5% level of significance.

Example :

Suppose we are interested in a population of 20 industrial units of the same size, all of which are experiencing excessive labour turnover problems. The past records show that the mean of the distribution of annual turnover is 320 employees, with a standard deviation of 75 employees. A sample of 5 of these industrial units is taken at random which gives a mean of annual turnover as 300 employees. Is the sample mean consistent with the population mean? Test at 5% level.

Solution: Taking the null hypothesis that the population mean is 320 employees, we can write:

$$H0: \mu_{H0} = 320 \text{ employees}$$

$$Ha: \mu_{H0} \neq 320 \text{ employees}$$

and the given information as under:

$\bar{X}$ = 300 employees, σ_p = 75 employees, n = 5; N = 20

Assuming the population to be normal, we can work out the test statistic *z* as under:

$$z^* = \frac{\bar{X} - \mu_{H_0}}{\sigma_p/\sqrt{n} \times \sqrt{(N-n)/(N-1)}}$$

$$= \frac{300 - 320}{75/\sqrt{5} \times \sqrt{(20-5)/(20-1)}} = -\frac{20}{(33.54)\,(.888)}$$

$$= -0.67$$

As *Ha* is two-sided in the given question, we shall apply a two-tailed test for determining the rejection regions at 5% level of significance which comes to as under, using normal curve area table:

$$R : |z| > 1.96$$

The observed value of z is –0.67 which is in the acceptance region since $R : |z| > 1.96$ and thus, $H0$ is accepted and we may conclude that the sample mean is consistent with population mean i.e., the population mean 320 is supported by sample results.

Example 3:

The mean of a certain production process is known to be 50 with a standard deviation of 2.5. The production manager may welcome any change is mean value towards higher side but would like to safeguard against decreasing values of mean. He takes a sample of 12 items that gives a mean value of 48.5. What inference should the manager take for the production process on the basis of sample results? Use 5 per cent level of significance for the purpose.

Solution: Taking the mean value of the population to be 50, we may write:

$H0$: $\mu_{H0} = 50$

Ha : $\mu_{H0} < 50$ (Since the manager wants to safeguard against decreasing values of mean.) and the given information as $\overline{X} = 48.5$, $\sigma_p = 2.5$ and $n = 12$. Assuming the population to be normal, we can work out the test statistic z as under:

$$z = \frac{\bar{X} - \mu_{H_0}}{\sigma_p/\sqrt{n}} = \frac{48.5 - 50}{2.5/\sqrt{12}} = -\frac{1.5}{(2.5)/(3.464)} = -2.0784$$

As *Ha* is one-sided in the given question, we shall determine the rejection region applying one tailed test (in the left tail because *Ha* is of less than type) at 5 per cent level of significance and it comes to as under, using normal curve area table:

$$R : z < -1.645$$

The observed value of *z* is – 2.0784 which is in the rejection region and thus, *H*0 is rejected at 5 per cent level of significance. We can conclude that the production process is showing mean which is significantly less than the population mean and this calls for some corrective action concerning the said process.

Example :

Raju Restaurant near the railway station at Falna has been having average sales of 500 tea cups per day. Because of the development of bus stand nearby, it expects to increase its sales. During the first 12 days after the start of the bus stand, the daily sales were as under:

550, 570, 490, 615, 505, 580, 570, 460, 600, 580, 530, 526

On the basis of this sample information, can one conclude that Raju Restaurant's sales have increased? Use 5 per cent level of significance.

Solution: Taking the null hypothesis that sales average 500 tea cups per day and they have not increased unless proved, we can write:

*H*0 : μ= 500 cups per day

Ha : μ> 500 (as we want to conclude that sales have increased).

As the sample size is small and the population standard deviation is not known, we shall use *t*-test assuming normal population and shall work out the test statistic *t* as:

$$t = \frac{\bar{X} - \mu}{\sigma_s/\sqrt{n}}$$

(To find $\bar{X}$ and σ_s we make the following computations:)

SL NO.	X	$X-\bar{X}$	$(X-\bar{X})^2$
1	550	2	4
2	570	22	484
3	490	-58	3364
4	615	67	4489
5	505	-43	1849
6	580	32	1024
7	570	22	484
8	460	-88	7744
9	600	52	2704
10	580	32	1024
11	530	-18	324
12	526	-22	484
Total	6576		23978
Mean	548		

$$\therefore \quad \bar{X} = \frac{\Sigma X_i}{n} = \frac{6576}{12} = 548$$

and
$$\sigma_s = \sqrt{\frac{\Sigma\left(X_i - \bar{X}\right)^2}{n-1}} = \sqrt{\frac{23978}{12-1}} = 46.68$$

Hence,
$$t = \frac{548 - 500}{46.68/\sqrt{12}} = \frac{48}{13.49} = 3.558$$

.

Degree of freedom = $n - 1 = 12 - 1 = 11$

As *Ha* is one-sided, we shall determine the rejection region applying one-tailed test (in the right tail because *Ha* is of more than type) at 5 per cent level of significance and it comes to as under, using table of *t*-distribution for 11 degrees of freedom:

$$R : t > 1.796$$

The observed value of *t* is 3.558 which is in the rejection region and thus *H*0 is rejected at 5 percent level of significance and we can conclude that the sample data indicate that Raju restaurant's sales have increased.

F-test:

When we want to test the equality of variances of two normal populations, we make use of *F*-test based on *F*-distribution. In such a situation, the null hypothesis happens to be $H_0: \sigma_{p_1}^2 = \sigma_{p_2}^2$, $\sigma_{p_1}^2$ and $\sigma_{p_2}^2$ representing the variances of two normal populations. This hypothesis is tested on the basis of sample data and the test statistic *F* is found, using $\sigma_{s_1}^2$ and $\sigma_{s_2}^2$ the sample estimates for $\sigma_{p_1}^2$ and $\sigma_{p_2}^2$ respectively, as stated below:

$$F = \frac{\sigma_{s_1}^2}{\sigma_{s_2}^2}$$

where $\sigma_{s_1}^2 = \frac{\Sigma\left(X_{1i} - \bar{X}_1\right)^2}{(n_1 - 1)}$ and $\sigma_{s_2}^2 = \frac{\Sigma\left(X_{2i} - \bar{X}_2\right)^2}{(n_2 - 1)}$

While calculating *F*, σ_{s1}^2 is treated > σ_{s2}^2 which means that the numerator is always the greater variance. Tables for *F*-distribution have been prepared by statisticians for different values of *F* at different levels of significance for different degrees of freedom for the greater and the smaller variances. By comparing the observed value of *F* with the corresponding table value, we can infer whether the difference between the variances of samples could have arisen due to sampling fluctuations. If the calculated value of *F* is greater than table value of *F* at a certain level of significance for (n_1 – 1) and (n_2 – 2) degrees of freedom, we regard the *F*-ratio as significant. Degrees of freedom for greater variance is represented as $v1$ and for smaller variance as v_2. On the other hand, if the calculated value of *F* is smaller than its table value, we conclude that *F*-ratio is not significant. If *F*-ratio is considered non-significant, we accept the null hypothesis, but if *F*-ratio is considered significant, we then reject H_0 (i.e., we accept H_a).

When we use the *F*-test, we presume that
(i) The populations are normal;
(ii) Samples have been drawn randomly;
(iii) Observations are independent; and
(iv) There is no measurement error.
The object of *F*-test is to test the hypothesis whether the two samples are from the same normal population with equal variance or from two normal populations with

equal variances. *F*-test was initially used to verify the hypothesis of equality between two variances, but is now mostly used in the context of analysis of variance. The following examples illustrate the use of *F*-test for testing the equality of variances of two normal populations.

Example

Two random samples drawn from two normal populations are:
Sample 1 20 16 26 27 23 22 18 24 25 19
Sample 2 27 33 42 35 32 34 38 28 41 43 30 37
Test using variance ratio at 5 per cent and 1 per cent level of significance whether the two populations have the same variances.

Solution: We take the null hypothesis that the two populations from where the samples have been drawn have the same variances i.e., $H_0 : \sigma_{p1}^2 = \sigma_{p2}^2$. From the sample data we work out σ_{s1}^2 and σ_{s2}^2 as under:

X_{1i}	$X_{1i} - \overline{X}_1$	$(X_{1i} - \overline{X}_1)^2$		X_{2i}	$X_{2i} - \overline{X}_2$	$(X_{2i} - \overline{X}_2)^2$
20	-2	4		27	-8	64
16	-6	36		33	-2	4
26	4	16		42	7	49
27	5	25		35	0	0
23	1	1		32	-3	9
22	0	0		34	-1	1
18	-4	16		38	3	9
24	2	4		28	-7	49
25	3	9		41	6	36
19	-3	9		43	8	64
				30	-5	25
				37	2	4
$\sum X_{1i}=220$		$\sum (X_{1i} - X_1)^2=120$		420		$\sum(X_{2i} - \overline{X}_2)^2=314$

$$\bar{X}_1 = \frac{\sum X_{1i}}{n_1} = \frac{220}{10} = 22; \qquad \bar{X}_2 = \frac{\sum X_{2i}}{n_2} = \frac{420}{12} = 35$$

$$\sigma^2_{s_1} = \frac{\sum \left(X_{1i} - \bar{X}_1\right)^2}{n_1 - 1} = \frac{120}{10 - 1} = 13.33$$

and $$\sigma^2_{s_2} = \frac{\sum \left(X_{2i} - \bar{X}_2\right)^2}{n_2 - 1} = \frac{314}{12 - 1} = 28.55$$

Hence, $$F = \frac{\sigma^2_{s_2}}{\sigma^2_{s_1}} \qquad \left(\because \sigma^2_{s_2} > \sigma^2_{s_1}\right)$$

$$= \frac{28.55}{13.33} = 2.14$$

Degrees of freedom in sample 1 = $(n_1 - 1) = 10 - 1 = 9$
Degrees of freedom in sample 2 = $(n_2 - 1) = 12 - 1 = 11$
As the variance of sample 2 is greater variance, hence

$$v_1 = 11;\ v_2 = 9$$

The table value of F at 5 per cent level of significance for $v_1 = 11$ and $v_2 = 9$ is 3.11 and the table value of F at 1 per cent level of significance for $v_1 = 11$ and $v_2 = 9$ is 5.20. Since the calculated value of $F = 2.14$ which is less than 3.11 and also less than 5.20, the F ratio is insignificant at 5 per cent as well as at 1 per cent level of significance and as such we accept the null hypothesis and conclude that samples have been drawn from two populations having the same variances.

Chi- square test:

The chi-square test is an important test amongst the several tests of significance developed by statisticians. Chi-square, symbolically written as $\chi 2$ (Pronounced as Ki-square), is a statistical measure used in the context of sampling analysis for comparing a variance to a theoretical variance.
As a non-parametric test, it "can be used to determine if categorical data shows dependency or the two classifications are independent. It can also be used to make

comparisons between theoretical populations and actual data when categories are used."

Thus, the chi-square test is applicable in large number of problems. The test is, in fact, a technique through the use of which it is possible for all researchers to

(i) test the goodness of fit;
(ii) test the significance of association between two attributes, and
(iii) test the homogeneity or the significance of population variance.

Chi-Square as a Test for Comparing Variance

The chi-square value is often used to judge the significance of population variance i.e., we can use the test to judge if a random sample has been drawn from a normal population with mean (μ) and with a specified variance (σ_p^2). The test is based on $\chi 2$ -distribution. Such a distribution we encounter when we deal with collections of values that involve adding up squares. Variances of samples require us to add a collection of squared quantities and, thus, have distributions that are related to $\chi 2$ -distribution. If we take each one of a collection of sample variances, divided them by the known population variance and multiply these quotients by (n – 1), where n means the number of items in the sample, we shall obtain a $\chi 2$ -distribution.

Thus,

$$\frac{\sigma_s^2}{\sigma_p^2}(n-1) = \frac{\sigma_s^2}{\sigma_p^2}\ (\text{d.f.})$$

would have the same distribution as $\chi 2$ -distribution with (n – 1) degrees of freedom.

In brief, when we have to use chi-square as a test of population variance, we have to work out the value of $\chi 2$ to test the null hypothesis (viz., H_0: $\sigma_s{}^2 = \sigma_p^2$) as under:

$$\chi^2 = \frac{\sigma_s^2}{\sigma_p^2}(n-1)$$

where σ_s^2 = variance of the sample;

σ_p^2 = variance of the population;

(n – 1) = degrees of freedom, n being the number of items in the sample.

Then by comparing the calculated value with the table value of $\chi 2$ for (n – 1) degrees of freedom at a given level of significance, we may either accept or reject the null hypothesis. If the calculated value of $\chi 2$ is less than the table value, the null hypothesis is accepted, but if the calculated value is equal or greater than the table value, the hypothesis is rejected.

Example :

A sample of 10 is drawn randomly from a certain population. The sum of the squared deviations from the mean of the given sample is 50. Test the hypothesis that the variance of the population is 5 at 5 per cent level of significance.

Solution: Given information is

$$n = 10$$

$$\Sigma\left(X_i - \bar{X}\right)^2 = 50$$

$$\therefore \quad \sigma_s^2 = \frac{\Sigma\left(X_i - \bar{X}\right)^2}{n-1} = \frac{50}{9}$$

Take the null hypothesis as $H_0 : \sigma_p^2 = \sigma_s^2$. In order to test this hypothesis, we work out the χ^2 value as under:

$$\chi^2 = \frac{\sigma_s^2}{\sigma_p^2}(n-1) = \frac{\frac{50}{9}}{5}(10-1) = \frac{50}{9} \times \frac{1}{5} \times \frac{9}{1} = 10$$

Degrees of freedom = (10 – 1) = 9.

The table value of χ^2 at 5 per cent level for 9 d.f. is 16.92. The calculated value of χ^2 is less than this table value, so we accept the null hypothesis and conclude that the variance of the population is 5 as given in the question.

CHI-SQUARE AS A NON-PARAMETRIC TEST

Chi-square is an important non-parametric test and as such no rigid assumptions are necessary in respect of the type of population. We require only the degrees of freedom (implicitly of course the size of the sample) for using this test. As a non-parametric test, chi-square can be used

(i) as a test of goodness of fit and

(ii) As a test of independence.

i. ***As a test of goodness of fit,*** $\chi 2$ test enables us to see how well does the assumed theoretical distribution (such as Binomial distribution, Poisson distribution or Normal distribution) fit to the observed data. When some theoretical distribution is fitted to the given data, we are always interested in knowing as to how well this distribution fits with the observed data. The chi-square test can give answer to this. If the calculated value of $\chi 2$ is less than the table value at a certain level of significance, the fit is considered to be a good one which means that the divergence between the observed and expected frequencies is attributable to fluctuations of sampling. But if the calculated value of $\chi 2$ is greater than its table value, the fit is not considered to be a good one.

ii. ***As a test of independence***, $\chi 2$ test enables us to explain whether or not two attributes are associated. For instance, we may be interested in knowing whether a new medicine is effective in controlling fever or not, $\chi 2$ test will helps us in deciding this issue. In such a situation, we proceed with the null hypothesis that the two attributes (viz., new medicine and control of fever) are independent which means that new medicine is not effective in controlling fever. On this basis we first calculate the expected frequencies and then work out the value of $\chi 2$. If the calculated value of $\chi 2$ is less than the table value at a certain level of significance for given degrees of freedom, we conclude that null hypothesis stands which means that the two attributes are independent or not associated (i.e., the new medicine is not effective in controlling the fever). But if the calculated value of $\chi 2$ is greater than its table value, our inference then would be that null hypothesis does not hold good which means the two attributes are associated and the association is not because of some chance factor but it exists in reality (i.e.,

the new medicine is effective in controlling the fever and as such may be prescribed). It may, however, be stated here that χ^2 is not a measure of the degree of relationship or the form of relationship between two attributes, but is simply a technique of judging the significance of such association or relationship between two attributes.

In order that we may apply the chi-square test either as a test of goodness of fit or as a test to judge the significance of association between attributes, it is necessary that the observed as well as theoretical or expected frequencies must be grouped in the same way and the theoretical distribution must be adjusted to give the same total frequency as we find in case of observed distribution. χ 2 is then calculated as follows:

$$\chi^2 = \Sigma \frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}}$$

where

O_{ij} = observed frequency of the cell in *i*th row and *j*th column.

E_{ij} = expected frequency of the cell in *i*th row and *j*th column.

CONDITIONS FOR THE APPLICATION OF χ^2 TEST

The following conditions should be satisfied before χ^2 test can be applied:

(i) Observations recorded and used are collected on a random basis,

(ii) All the items in the sample must be independent.

(iii) No group should contain very few items, say less than 10. In case where the frequencies are less than 10, regrouping is done by combining the frequencies of adjoining groups so that the new frequencies become greater than 10. Some statisticians take this number as 5, but 10 is regarded as better by most of the statisticians.

(iv) The overall number of items must also be reasonably large. It should normally be at least

50, howsoever small the number of groups may be.

(v) The constraints must be linear. Constraints which involve linear equations in the cell frequencies of a contingency table (i.e., equations containing no squares or higher powers of the frequencies) are known as linear constraints.

STEPS INVOLVED IN APPLYING CHI-SQUARE TEST

The various steps involved are as follows:

(i) First of all calculate the expected frequencies on the basis of given hypothesis or on the basis of null hypothesis. Usually in case of a 2×2 or any contingency table, the expected frequency for any given cell is worked out as under:

$$\text{Expected frequency of any cell} = \left[\frac{(\text{Row total for the row of that cell}) \times (\text{Column total for the column of that cell})}{(\text{Grand total})}\right]$$

(ii) Obtain the difference between observed and expected frequencies and find out the squares of such differences i.e., calculate $(O_{ij} - E_{ij})^2$.

(iii) Divide the quantity $(O_{ij} - E_{ij})^2$ obtained as stated above by the corresponding expected frequency to get $(O_{ij} - E_{ij})^2/E_{ij}$ and this should be done for all the cell frequencies or the group frequencies.

(iv) Find the summation of $(O_{ij} - E_{ij})^2/E_{ij}$ values or what we call $\sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$. This is the required χ^2 value.

The χ^2 value obtained as such should be compared with relevant table value of χ^2 and then inference be drawn as stated above.

Example:

Find the value of $\chi 2$ for the following information:

Class	A	B	C	D	E
Observed Frequency	8	29	44	15	4
Theoretical (Or Expected) Frequency	7	24	38	24	7

Solution: Since some of the frequencies less than 10, we shall first re-group the given data as follows and then will work out the value of $\chi 2$:

CLASS	OBSERVED FREQUENCY Oi	EXPECTED FREQUENCY Ei	Oi - Ei	(Oi - Ei)2/Ei
A and B	8 + 29 = 37	7 + 24 = 31	6	36/31
C	44	38	6	36/38
D and E	15 + 4 = 19	24 + 7 = 31	-12	144/31

$$\therefore \qquad \chi^2 = \Sigma \frac{(O_i - E_i)^2}{E_i} = 6.76 \text{ app.}$$

Example

The table given below shows the data obtained during outbreak of

	ATTACKED	NOT ATTACKED	TOTAL
VACCINATED	31	469	500
NOT VACCINATED	185	1315	1500
TOTAL	216	1784	2000

Test the effectiveness of vaccination in preventing the attack from smallpox. Test your result with the help of χ^2 at 5 per cent level of significance.

Solution: Let us take the hypothesis that vaccination is not effective in preventing the attack from smallpox i.e., vaccination and attack are independent. On the basis of this hypothesis, the expected frequency corresponding to the number of persons vaccinated and attacked would be:

Expectation of (AB) = {(A) X (B)} / N

when A represents vaccination and B represents attack.

$(A) = 500$
$(B) = 216$
$N = 2000$

Expectation of $(AB) = (500 * 216) / 2000 = 54$
Now using the expectation of (AB), we can write the table of expected values as follows:

	Attacked : B	Not Attacked : b	Total
Vaccinated : A	(AB) = 54	(Ab) = 446	500
Not Vaccinated : a	(aB) = 162	(ab) = 1338	1500
Total	216	1784	2000

GROUP	OBSERVED FREQUENCY Oi	EXPECTED FREQUENCY Ei	Oi - Ei	(Oi - Ei)2/Ei
AB	31	54	-23	529/54
Ab	469	446	23	529/446
aB	185	162	23	529/162
ab	1315	1338	-23	529/1338

$$\chi^2 = \sum \frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}} = 14.642$$

Degrees of freedom in this case = $(r - 1)(c - 1) = (2 - 1)(2 - 1) = 1$.
The table value of $\chi 2$ for 1 degree of freedom at 5 per cent level of significance is 3.841. The calculated value of $\chi 2$ is much higher than this table value and hence the result of the experiment does not support the hypothesis. We can, thus, conclude that vaccination is effective in preventing the attack from smallpox.

Example

Two research workers classified some people in income groups on the basis of sampling studies. Their results are as follows:

Investigators	income groups			Total
	Poor	Middle	Rich	
A	160	30	10	200
B	140	120	40	300
	300	150	50	500

Show that the sampling technique of at least one research worker is defective.

Solution: Let us take the hypothesis that the sampling techniques adopted by research workers are similar (i.e., there is no difference between the techniques adopted by research workers). This being so, the expectation of *A* investigator classifying the people in

(i) Poor income group $= \frac{200 \times 300}{500} = 120$

(ii) Middle income group $= \frac{200 \times 150}{500} = 60$

(iii) Rich income group $= \frac{200 \times 50}{500} = 20$

Similarly the expectation of *B* investigator classifying the people in

(i) Poor income group $= \frac{300 \times 300}{500} = 180$

(ii) Middle income group $= \frac{300 \times 150}{500} = 90$

(iii) Rich income group $= \frac{300 \times 50}{500} = 30$

We can now calculate value of $\chi 2$ as follows:

Groups	Observed frequency Oi	Expected frquency Ei	Oi - Ei	(Oi - Ei)2/Ei
Investigators A				
Classifies People as poor	160	120	40	13.33333333
Classifies People as middle	30	60	-30	15
Classifies People as Rich	10	20	-10	5

Investigators B				
Classifies People as poor	140	180	-40	8.888888889
Classifies People as middle	120	90	30	10
Classifies People as Rich	40	30	10	3.333333333
				55.55555556

degree of freedom	(c-1)(r-1) (3-1)*(2-1) =	2

tabulated value at 5%level of significance	5.991

The table value of $\chi 2$ for two degrees of freedom at 5 per cent level of significance is 5.991.

The calculated value of $\chi 2$ is much higher than this table value which means that the calculated value cannot be said to have arisen just because of chance. It is significant. Hence, the hypothesis does not hold good. This means that the sampling techniques adopted by two investigators differ and are not similar. Naturally, then the technique of one must be superior to that of the other.

ANOVA TEST:

An extremely useful technique concerning researches in the fields of economics, biology, education, psychology, sociology, business/industry and in researches of several other disciplines. This technique is used when multiple sample cases are involved. As stated earlier, the significance of the difference between the means of two samples can be judged through either *z*-test or the *t*-test, but the difficulty arises when we happen to examine the significance of the difference amongst more than two sample means at the same time. The ANOVA technique enables us to

perform this simultaneous test and as such is considered to be an important tool of analysis in the hands of a researcher. Using this technique, one can draw inferences about whether the samples have been drawn from populations having the same mean.

THE BASIC PRINCIPLE OF ANOVA

The basic principle of ANOVA is to test for differences among the means of the populations by examining the amount of variation within each of these samples, relative to the amount of variation between the samples.

$$F = \frac{\text{Estimate of population variance based on between samples variance}}{\text{Estimate of population variance based on within samples variance}}$$

ONE WAY ANOVA:

One-way (*or single factor*) *ANOVA:* Under the one-way ANOVA, we consider only one factor and then observe that the reason for said factor to be important is that several possible types of samples can occur within that factor. We then determine if there are differences within that factor.

The technique involves the following steps:

(i) Obtain the mean of each sample i.e., obtain

$$\bar{X}_1, \bar{X}_2, \bar{X}_3, \ldots, \bar{X}_k$$

when there are *k* samples.

(ii) Work out the mean of the sample means as follows:

$$\bar{\bar{X}} = \frac{\bar{X}_1 + \bar{X}_2 + \bar{X}_3 + \ldots + \bar{X}_k}{\text{No. of samples } (k)}$$

(iii) Take the deviations of the sample means from the mean of the sample means and calculate the square of such deviations which may be multiplied by the number of items in the corresponding sample, and then obtain their total. This is known as

the sum of squares for variance between the samples (or *SS* between). Symbolically, this can be written:

$$SS \text{ between} = n_1\left(\bar{X}_1 - \bar{\bar{X}}\right)^2 + n_2\left(\bar{X}_2 - \bar{\bar{X}}\right)^2 + \ldots + n_k\left(\bar{X}_k - \bar{\bar{X}}\right)^2$$

(iv) Divide the result of the (iii) step by the degrees of freedom between the samples to obtain variance or mean square (*MS*) between samples. Symbolically, this can be written:

$$MS \text{ between} = \frac{SS \text{ between}}{(k-1)}$$

(k – 1) represents degrees of freedom (d.f.) between samples.

(v) Obtain the deviations of the values of the sample items for all the samples from corresponding means of the samples and calculate the squares of such deviations and then obtain their total. This total is known as the sum of squares for variance within samples (or *SS* within).
Symbolically this can be written:

$$SS \text{ within} = \Sigma\left(X_{1i} - \bar{X}_1\right)^2 + \Sigma\left(X_{2i} - \bar{X}_2\right)^2 + \ldots + \Sigma\left(X_{ki} - \bar{X}_k\right)^2$$
$$i = 1, 2, 3, \ldots$$

(vi) Divide the result of (v) step by the degrees of freedom within samples to obtain the variance or mean square (*MS*) within samples. Symbolically, this can be written:

$$MS \text{ within} = \frac{SS \text{ within}}{(n-k)}$$

where $(n - k)$ represents degrees of freedom within samples,
n = total number of items in all the samples i.e., $n_1 + n_2 + \ldots + n_k$
k = number of samples.

(vii) For a check, the sum of squares of deviations for total variance can also be worked out by adding the squares of deviations when the deviations for the individual items in all the samples have been taken from the mean of the sample means. Symbolically, this can be written:

$$SS \text{ for total variance } = \sum\left(X_{ij} - \overline{\overline{X}}\right)^2 \qquad i = 1, 2, 3, \ldots$$

$$j = 1, 2, 3, \ldots$$

This total should be equal to the total of the result of the (iii) and (v) steps explained above

i.e., *SS* for total variance = *SS* between + *SS* within.

The degrees of freedom for total variance will be equal to the number of items in all samples minus one i.e., (n – 1). The degrees of freedom for between and within must add up to the degrees of freedom for total variance i.e., (n – 1) = (k – 1) + (n – k)

This fact explains the additive property of the ANOVA technique.

(viii) Finally, *F*-ratio may be worked out as under:

$$F\text{-ratio} = \frac{MS \text{ between}}{MS \text{ within}}$$

This ratio is used to judge whether the difference among several sample means is significant or is just a matter of sampling fluctuations. For this purpose we look into the table, giving the values of *F* for given degrees of freedom at different levels of significance. If the worked out value of *F*, as stated above, is less than the table value of *F*, the difference is taken as insignificant i.e., due to chance and the null-hypothesis of no difference between sample means stands. In case the calculated value of *F* happens to be either equal or more than its table value, the difference is considered as significant (which means the samples could not have come from the same universe) and accordingly the conclusion may be drawn. The higher the calculated value of *F* is above the table value, the more definite and sure one can be about his conclusions.

Source of variation	*Sum of squares (SS)*	*Degrees of freedom (d.f.)*	*Mean Square (MS) (This is SS divided by d.f.) and is an estimation of variance to be used in F-ratio*	*F-ratio*
Between samples or categories	$n_1\left(\overline{X}_1 - \overline{\overline{X}}\right)^2 + \ldots + n_k\left(\overline{X}_k - \overline{\overline{X}}\right)^2$	$(k-1)$	$\frac{SS \text{ between}}{(k-1)}$	$\frac{MS \text{ between}}{MS \text{ within}}$
Within samples or categories	$\Sigma\left(X_{1i} - \overline{X}_1\right)^2 + \ldots + \Sigma\left(X_{ki} - \overline{X}_k\right)^2$ $i = 1, 2, 3, \ldots$	$(n-k)$	$\frac{SS \text{ within}}{(n-k)}$	
Total	$\Sigma\left(X_{ij} - \overline{\overline{X}}\right)^2$ $i = 1, 2, \ldots$ $j = 1, 2, \ldots$	$(n-1)$		

Example

Set up an analysis of variance table for the following per acre production data for three varieties of wheat, each grown on 4 plots and state if the variety differences are significant.

Plot Of Land	Per Acre Production Data		
	Variety Of Wheat		
	A	B	C
1	6	5	5
2	7	5	4
3	3	3	3
4	8	7	4

Solution: We can solve the problem by the direct method or by short-cut method, but in each case we shall get the same result. We try below both the methods.

Solution through direct method: First we calculate the mean of each of these samples:

$$\bar{X}_1 = \frac{6+7+3+8}{4} = 6$$

$$\bar{X}_2 = \frac{5+5+3+7}{4} = 5$$

$$\bar{X}_3 = \frac{5+4+3+4}{4} = 4$$

Mean of the sample means or $\quad \bar{\bar{X}} = \dfrac{\bar{X}_1 + \bar{X}_2 + \bar{X}_3}{k}$

$$= \frac{6+5+4}{3} = 5$$

Now we work out *SS* between and *SS* within samples:

$$SS \text{ between} = n_1\left(\bar{X}_1 - \bar{\bar{X}}\right)^2 + n_2\left(\bar{X}_2 - \bar{\bar{X}}\right)^2 + n_3\left(\bar{X}_3 - \bar{\bar{X}}\right)^2$$

$$= 4(6-5)^2 + 4(5-5)^2 + 4(4-5)^2$$
$$= 4 + 0 + 4$$
$$= 8$$

$$SS \text{ within} = \Sigma\left(X_{1i} - \bar{X}_1\right)^2 + \Sigma\left(X_{2i} - \bar{X}_2\right)^2 + \Sigma\left(X_{3i} - \bar{X}_3\right)^2, \qquad i = 1, 2, 3, 4$$

$$= \{(6-6)^2 + (7-6)^2 + (3-6)^2 + (8-6)^2\}$$
$$+ \{(5-5)^2 + (5-5)^2 + (3-5)^2 + (7-5)^2\}$$
$$+ \{(5-4)^2 + (4-4)^2 + (3-4)^2 + (4-4)^2\}$$
$$= \{0+1+9+4\} + \{0+0+4+4\} + \{1+0+1+0\}$$
$$= 14 + 8 + 2$$
$$= 24$$

$$SS \text{ for total variance } = \sum\left(X_{ij} - \overline{\overline{X}}\right)^2 \qquad i = 1, 2, 3\ldots$$

$$j = 1, 2, 3\ldots$$

$$= (6-5)^2 + (7-5)^2 + (3-5)^2 + (8-5)^2 + (5-5)^2 + (5-5)^2 + (3-5)^2 + (7-5)^2 + (5-5)^2 + (4-5)^2 + (3-5)^2 + (4-5)^2$$

$$= 1 + 4 + 4 + 9 + 0 + 0 + 4 + 4 + 0 + 1 + 4 + 1$$

$$= 32$$

Alternatively, it (*SS* for total variance) can also be worked out thus:

SS for total = *SS* between + *SS* within

= 8 + 24

= 32

We can now set up the ANOVA table for this problem:

Sources Of Variation	SS	d.f.	MS	F-ratio	5% F-limit from F-table
Between Sample	8	(3-1)=2	8/2=4	4/2.67=1.5	F(2,9)= 4.26
Within Sample	24	(12-3)=9	24/9=2.67		
Total	32	11			

The above table shows that the calculated value of *F* is 1.5 which is less than the table value of 4.26 at 5% level with d.f. being $v1 = 2$ and $v2 = 9$ and hence could have arisen due to chance. This analysis supports the null-hypothesis of no difference is sample means. We may, therefore, conclude that the difference in wheat output due to varieties is insignificant and is just a matter of chance.

TWO-WAY ANOVA:

Two-way ANOVA technique is used when the data are classified on the basis of two factors. For example, the agricultural output may be classified on the basis of

different varieties of seeds and also on the basis of different varieties of fertilizers used. A business firm may have its sales data classified on the basis of different salesmen and also on the basis of sales in different regions. In a factory, the various units of a product produced during a certain period may be classified on the basis of different varieties of machines used and also on the basis of different grades of labour. Such a two-way design may have repeated measurements of each factor or may not have repeated values. The ANOVA technique is little different in case of repeated measurements where we also compute the interaction variation. We shall now explain the two-way ANOVA technique in the context of both the said designs with the help of examples.

(a) *ANOVA technique in context of two-way design when repeated values are not there:* As we do not have repeated values, we cannot directly compute the sum of squares within samples as we had done in the case of one-way ANOVA. Therefore, we have to calculate this residual or error variation by subtraction, once we have calculated (just on the same lines as we did in the case of one way ANOVA) the sum of squares for total variance and for variance between varieties of one treatment as also for variance between varieties of the other treatment.

The various steps involved are as follows:

(i) Use the coding device, if the same simplifies the task.
(ii) Take the total of the values of individual items (or their coded values as the case may be) in all the samples and call it *T*.
(iii) Work out the correction factor as under:

$$\text{Correction factor} = \frac{(T)^2}{n}$$

(iv) Find out the square of all the item values (or their coded values as the case may be) one by one and then take its total. Subtract the correction factor from this total to obtain the sum of squares of deviations for total variance. Symbolically, we can write it as:

Sum of squares of deviations for total variance or total *SS*

$$= \sum X_{ij}^2 - \frac{(T)^2}{n}$$

(v) Take the total of different columns and then obtain the square of each column total and divide such squared values of each column by the number of items in the concerning column and take the total of the result thus obtained. Finally, subtract the correction factor from this total to obtain the sum of squares of deviations for variance between columns or (*SS* between columns).

(vi) Take the total of different rows and then obtain the square of each row total and divide such squared values of each row by the number of items in the corresponding row and take the total of the result thus obtained. Finally, subtract the correction factor from this total to obtain the sum of squares of deviations for variance between rows (or *SS* between rows).

(vii) Sum of squares of deviations for residual or error variance can be worked out by subtracting the result of the sum of (v)th and (vi)th steps from the result of (iv) th step stated above. In other words,

Total *SS* – (*SS* between columns + *SS* between rows)

= *SS* for residual or error variance.

(viii) Degrees of freedom (d.f.) can be worked out as under:

d.f. for total variance = (c . r – 1)

d.f. for variance between columns = (c – 1)

d.f. for variance between rows = (r – 1)

d.f. for residual variance = (c – 1) (r – 1)

where c = number of columns

r = number of rows

(ix) ANOVA table can be set up in the usual fashion as shown below:

Source of variation	*Sum of squares (SS)*	*Degrees of freedom (d.f.)*	*Mean square (MS)*	*F-ratio*
Between columns treatment	$\sum \frac{(T_j)^2}{n_j} - \frac{(T)^2}{n}$	$(c-1)$	$\frac{SS \text{ between columns}}{(c-1)}$	$\frac{MS \text{ between columns}}{MS \text{ residual}}$
Between rows treatment	$\sum \frac{(T_i)^2}{n_i} - \frac{(T)^2}{n}$	$(r-1)$	$\frac{SS \text{ between rows}}{(r-1)}$	$\frac{MS \text{ between rows}}{MS \text{ residual}}$
Residual or error	Total SS – (SS between columns + SS between rows)	$(c-1)(r-1)$	$\frac{SS \text{ residual}}{(c-1)(r-1)}$	
Total	$\sum X_{ij}^2 - \frac{(T)^2}{n}$	$(c.r-1)$		

In the table c = number of columns

r = number of rows

SS residual = Total SS – (SS between columns + SS between rows).

Thus, MS residual or the residual variance provides the basis for the F-ratios concerning variation between columns treatment and between rows treatment. MS residual is always due to the fluctuations of sampling, and hence serves as the basis for the significance test. Both the F-ratios are compared with their corresponding table values, for given degrees of freedom at a specified level of significance, as usual and if it is found that the calculated F-ratio concerning variation between columns is equal to or greater than its table value, then the difference among columns means is considered significant. Similarly, the F-ratio concerning variation between rows can be interpreted.

Example

Set up an analysis of variance table for the following two-way design results:

Per Acre Production Data Of Wheat			
Varieties Of Seeds	A	B	C
Varieties Of Fertilizers			
1	6	5	5
2	7	5	4
3	3	3	3
4	8	7	4

Also state whether variety differences are significant at 5% level.

Solution:

As the given problem is a two-way design of experiment without repeated values, we shall adopt all the above stated steps while setting up the ANOVA table

Step (i) $T = 60, n = 12, \therefore$ Correction factor $= \dfrac{(T)^2}{n} = \dfrac{60 \times 60}{12} = 300$

Step (ii) Total $SS = (36 + 25 + 25 + 49 + 25 + 16 + 9 + 9 + 9 + 64 + 49 + 16) - \left(\dfrac{60 \times 60}{12}\right)$

$= 332 - 300$

$= 32$

Step (iii) SS between columns treatment $= \left[\dfrac{24 \times 24}{4} + \dfrac{20 \times 20}{4} + \dfrac{16 \times 16}{4}\right] - \left[\dfrac{60 \times 60}{12}\right]$

$= 144 + 100 + 64 - 300$

$= 8$

Step (iv) SS between rows treatment $= \left[\dfrac{16 \times 16}{3} + \dfrac{16 \times 16}{3} + \dfrac{9 \times 9}{3} + \dfrac{19 \times 19}{3}\right] - \left[\dfrac{60 \times 60}{12}\right]$

$= 85.33 + 85.33 + 27.00 + 120.33 - 300$

$= 18$

Step (v) SS residual or error = Total $SS - (SS$ between columns + SS between rows)

$= 32 - (8 + 18)$

$= 6$

Source of variation	*SS*	*d.f.*	*MS*	*F-ratio*	*5% F-limit (or the tables values)*
Between columns (i.e., between varieties of seeds)	8	$(3-1)=2$	$8/2=4$	$4/1=4$	$F(2,6)=5.14$
Between rows (i.e., between varieties of fertilizers)	18	$(4-1)=3$	$18/3=6$	$6/1=6$	$F(3,6)=4.76$
Residual or error	6	$(3-1)\times(4-1)=6$	6/6=1		
Total	32	$(3\times4)-1=11$			

From the said ANOVA table, we find that differences concerning varieties of seeds are insignificant at 5% level as the From calculated *F*-ratio of 4 is less than the table value of 5.14, but the variety differences concerning fertilizers are significant as the calculated *F*-ratio of 6 is more than its table value of 4.76.

NON PARAMETRIC TEST

SIGN TEST:

The sign test is one of the easiest parametric tests. Its name comes from the fact that it is based on the direction of the plus or minus signs of observations in a sample and not on their numerical magnitudes. The sign test may be one of the following two types:

(a) One sample sign test;

(b) Two sample sign test.

(a) ***One sample sign test:*** The one sample sign test is a very simple non-parametric test applicable when we sample a continuous symmetrical population in which case the probability of getting a sample value less than mean is 1/2 and the probability of getting a sample value greater than mean is also 1/2. To test the null hypothesis $\mu = \mu_{H0}$ against an appropriate alternative on the basis of a random sample of size '*n*', we replace the value of each and every item of the sample with a plus (+) sign if it is greater than μ_{H0}, and with a minus (–) sign if it is less than μ_{H0} . But if the value happens

to be equal to μ_{H0}, then we simply discard it. After doing this, we test the null hypothesis that these + and – signs are values of a random variable, having a binomial distribution with $p = 1/2$. For performing one sample sign test when the sample is small, we can use tables of binomial probabilities, but when sample happens to be large, we use normal approximation to binomial distribution. Let us take an illustration to apply one sample sign test.

(b) ***Two sample sign test* (*or the sign test for paired data*):** The sign test has important applications in problems where we deal with paired data. In such problems, each pair of values can be replaced with a plus (+) sign if the first value of the first sample (say *X)* is greater than the first value of the second sample (say *Y*) and we take minus (–) sign if the first value of *X* is less than the first value of *Y*. In case the two values are equal, the concerning pair is discarded. (In case the two samples are not of equal size, then some of the values of the larger sample left over after the random pairing will have to be discarded.) The testing technique remains the same as started in case of one sample sign test. An example can be taken to explain and illustrate the two sample sign test.

Example

Suppose playing four rounds of golf at the City Club 11 professionals totalled 280, 282, 290, 273, 283, 283, 275, 284, 282, 279, and 281. Use the sign test at 5% level of significance to test the null hypothesis that professional golfers average $\mu_{H0} = 284$ for four rounds against the alternative hypothesis $\mu_{H0} < 284$.

Solution: To test the null hypothesis $\mu_{H0} = 284$ against the alternative hypothesis $\mu_{H0} < 284$ at 5% (or 0.05) level of significance, we first replace each value greater than 284 with a plus sign and each value less than 284 with a minus sign and discard the one value which actually equals 284. If we do this we get

–,–,+,–,–,–,–,–,–,–.

Now we can examine whether the one plus sign observed in 10 trials support the null hypothesis $p = 1/2$ or the alternative hypothesis $p < 1/2$. The probability of one or fewer successes with $n = 10$ and $p = 1/2$ can be worked out as under:

$$^{10}C_1 p^1 q^9 + {}^{10}C_0 p^0 q^{10} = 10\left(\frac{1}{2}\right)^1\left(\frac{1}{2}\right)^9 + 1\left(\frac{1}{2}\right)^0\left(\frac{1}{2}\right)^{10}$$

$$= 0.010 + 0.001$$

Values can also be seen from the table of binomial probabilities when $p = 1/2$ and $n = 10) = 0.011$

Since this value is less than $\alpha = 0.05$, the null hypothesis must be rejected. In other words, we conclude that professional golfers' average is less than 284 for four rounds of golf.

Alternatively, we can as well use normal approximation to the binomial distribution. If we do that, we find the observed proportion of success, on the basis of signs that we obtain, is 1/10 and that of failure is 9/10. The standard error of proportion assuming null hypothesis $p = 1/2$ is as under:

$$\sigma_{\text{prop.}} = \sqrt{\frac{p \cdot q}{n}} = \sqrt{\frac{\frac{1}{2} \times \frac{1}{2}}{10}} = 0.1581$$

For testing the null hypothesis i.e., $p = 1/2$ against the alternative hypothesis $p < 1/2$, a one-tailed test is appropriate which can be indicated as shown in the Fig. 12.1.

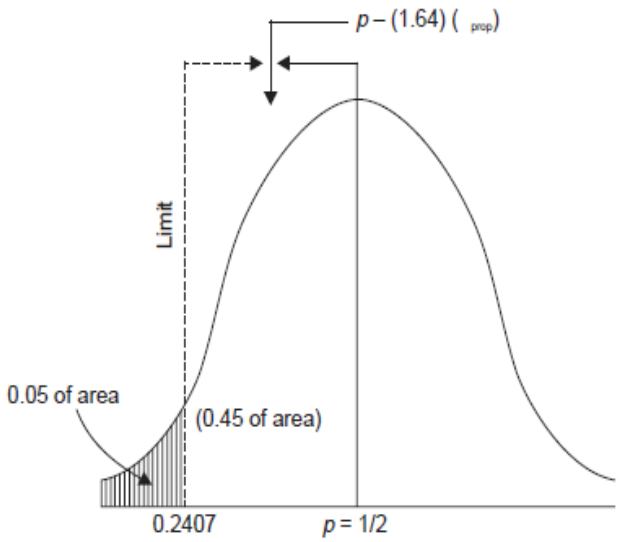
By using table of area under normal curve, we find the appropriate z value for 0.45 of the area under normal curve and it is 1.64. Using this, we now work out the limit (on the lower side as the alternative hypothesis is of < type) of the acceptance region as under:

$$p - z \cdot \sigma_{(\text{prop.})}$$

or $$p - (1.64)\,(0.1581)$$

or $$\frac{1}{2} - 0.2593$$

or $$0.2407$$



(Shaded portion indicates rejection region)

As the observed proportion of success is only 1/10 or 0.1 which comes in the rejection region, we reject the null hypothesis at 5% level of significance and accept the alternative hypothesis. Thus, we conclude that professional golfers' average is less than 284 for four rounds.

Example:

The following are the numbers of artifacts dug up by two archaeologists at an ancient cliff dwelling on 30 days

By X	1	0	2	3	1	0	2	2	3	0	1	1	4	1	2	1	3	5	2	1	3	2	4	1	3	2	0	2	4	2
By Y	0	0	1	0	2	0	0	1	1	2	0	1	2	1	1	0	2	2	6	0	2	3	0	2	1	0	1	0	1	0

Use the sign test at 1% level of significance to test the null hypothesis that the two archaeologists, *X* and *Y*, are equally good at finding artifacts against the alternative hypothesis that *X* is better.

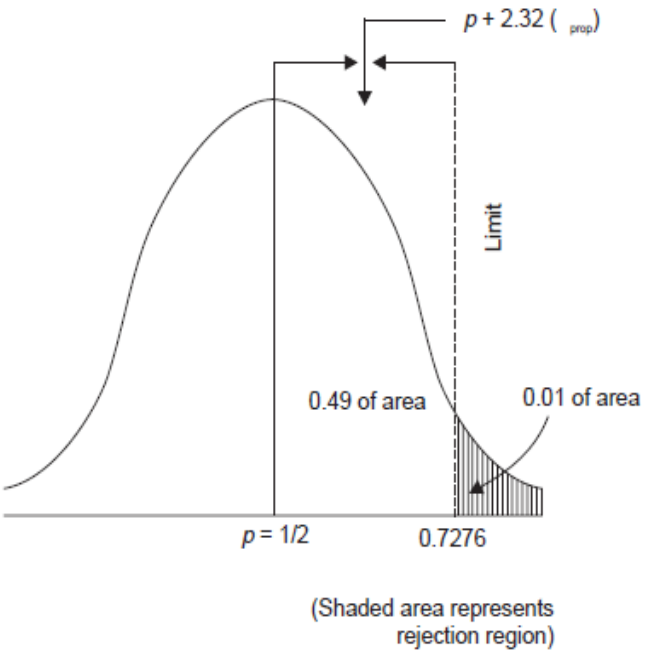
Solution: First of all the given paired values are changed into signs (+ or –) as under:

By X	1	0	2	3	1	0	2	2	3	0	1	1	4	1	2	1	3	5	2	1	3	2	4	1	3	2	0	2	4	2
By Y	0	0	1	0	2	0	0	1	1	2	0	1	2	1	1	0	2	2	6	0	2	3	0	2	1	0	1	0	1	0
Sign (X-Y)	+	0	+	+	-	0	+	+	+	-	+	0	+	0	+	+	+	+	-	+	+	-	+	-	+	+	-	+	+	+

Total Number of + signs = 20
Total Number of – signs = 6
Hence, sample size = 26
(Since there are 4 zeros in the sign row and as such four pairs are discarded, we are left with 30 – 4 = 26.)
Thus the observed proportion of pluses (or successes) in the sample is = 20/26 = 0.7692 and the observed proportion of minuses (or failures) in the sample is = 6/26 = 0.2308.
As we are to test the null hypothesis that the two archaeologists *X* and *Y* are equally good and if that is so, the number of pluses and minuses should be equal and as such $p = 1/2$ and $q = 1/2$. Hence, the standard error of proportion of successes, given the null hypothesis and the size of the sample, we have:

$$\sigma_{\text{prop.}} = \sqrt{\frac{p \cdot q}{n}} = \sqrt{\frac{\frac{1}{2} \times \frac{1}{2}}{26}} = 0.0981$$

Since the alternative hypothesis is that the archaeologists *X* is better (or $p > 1/2$), we find one tailed test is appropriate. This can be indicated as under, applying normal approximation to binomial distribution in the given case:



By using the table of area under normal curve, we find the appropriate *z* value for 0.49 of the area under normal curve and it is 2.32. Using this, we now work out the limit (on the upper side as the alternative hypothesis is of > type) of the acceptance region as under:

$$p + 2.32\sigma_{\text{prop.}} = 0.5 + 2.32(0.0981)$$
$$= 0.5 + 0.2276 = 0.7276$$

and we now find the observed proportion of successes is 0.7692 and this comes in the rejection region and as such we reject the null hypothesis, at 1% level of significance, that two archaeologists *X* and *Y* are equally good. In other words, we accept the alternative hypothesis, and thus conclude that archaeologist *X* is better.

Sign tests, as explained above, are quite simple and they can be applied in the context of both one-tailed and two-tailed tests. They are generally based on binomial distribution, but when the sample size happens to be large enough (such that $n \times p$ and $n \times q$ both happen to be greater than 5); we can as well make use of normal approximation to binomial distribution.

Runs Test:

One sample runs test is a test used to judge the randomness of a sample on the basis of the order in which the observations are taken. There are many applications in which it is difficult to decide whether the sample used is a random one or not. This is particularly true when we have little or no control over the selection of the data. For instance, if we want to predict a retail store's sales volume for a given month, we have no choice but to use past sales data and perhaps prevailing conditions in general. None of this information constitutes a random sample in the strict sense. To allow us to test samples for the randomness of their order, statisticians have developed the theory of runs. A run is a succession of identical letters (or other kinds of symbols) which is followed and preceded by different letters or no letters at all. To illustrate, we take the following arrangement of healthy, *H*, and diseased, *D*, mango trees that were planted many years ago along a certain road:

HH	*DD*	*HHHHH*	*DDD*	*HHHH*	*DDDDD*	*HHHHHHHHH*
1st	2nd	3rd	4th	5th	6th	7th

Using underlines to combine the letters which constitute the runs, we find that first there is a run of
two *H*'*s*, then a run of two *D*'*s*, then a run of five *H*'*s*, then a run of three *D*'*s*, then a run of four *H*'*s*, then a run of five *D*'*s* and finally a run of nine *H*'*s*. In this way there are 7 runs in all or $r = 7$. If there are too few runs, we might suspect a definite grouping or a trend; if there are too many runs, we might suspect some sort of repeated alternating patterns. In the given case there seems some grouping i.e., the diseased trees seem to come in groups. Through one sample runs test which is based on the idea that too few or too many runs show that the items were not chosen randomly, we can say whether the apparently seen grouping is significant or whether it can be attributed to chance. We shall use the following symbols for a test of runs:

n_1 = number of occurrences of type 1 (say *H* in the given case)
n_2 = number of occurrences of type 2 (say *D* in the given case)
r = number of runs.

In the given case the values of n_1, n_2 and r would be as follows:
$n_1 = 20$; $n_2 = 10$; $r = 7$

The sampling distribution of '*r*' statistic, the number of runs, is to be used and this distribution has its mean

$$\mu_r = \frac{2n_1n_2}{n_1 + n_2} + 1$$

and the standard deviation $\sigma_r = \sqrt{2n_1n_2 \frac{2n_1n_2 - n_1 - n_2}{(n_1 + n_2)^2 (n_1 + n_2 - 1)}}$

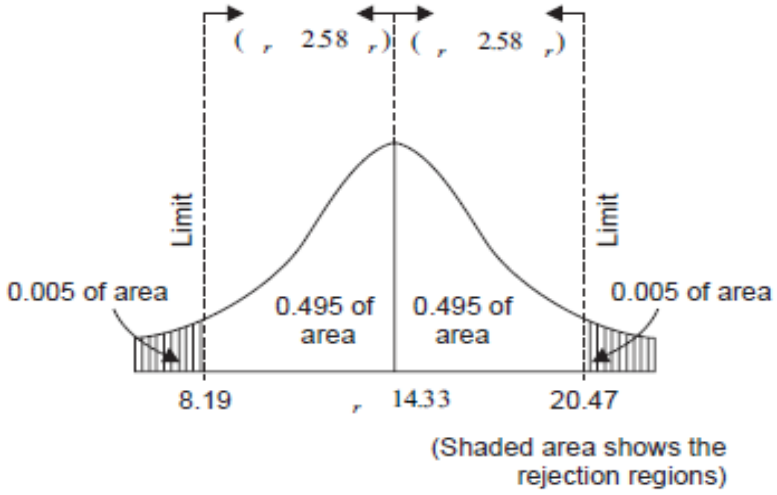
In the given case, we work out the values of μ_r and σ_r as follows:

$$\mu_r = \frac{(2)(20)(10)}{20 + 10} + 1 = 14.33$$

and

$$\sigma_r = \sqrt{\frac{(2)(20)(10)(2 \times 20 \times 10 - 20 - 10)}{(20 + 10)^2 (20 + 10 - 1)}} = 2.38$$

For testing the null hypothesis concerning the randomness of the planted trees, we should have been given the level of significance. Suppose it is 1% or 0.01. Since too many or too few runs would indicate that the process by which the trees were planted was not random, a two-tailed test is appropriate which can be indicated as follows on the assumption that the sampling distribution of *r* can be closely approximated by the normal distribution.



By using the table of area under normal curve, we find the appropriate *z* value for 0.495 of the area under the curve and it is 2.58. Using this we now calculate the limits of the acceptance region:

Upper limit = μ_r + (2.58) (2.38) = 14.33 + 6.14 = 20.47 and
Lower limit = μ_r – (2.58) (2.38) = 14.33 – 6.14 = 8.19
We now find that the observed number of runs (i.e., r = 7) lies outside the acceptance region i.e., in the rejection region. Therefore, we cannot accept the null hypothesis of randomness at the given level of significance viz., α = 0.01. As such we conclude that there is a strong indication that the diseased trees come in non-random grouping.
One sample runs test, as explained above, is not limited only to test the randomness of series of attributes. Even a sample consisting of numerical values can be treated similarly by using the letters say '*a*' and '*b*' to denote respectively the values falling above and below the median of the sample.
Numbers equal to the median are omitted. The resulting series of *a*'s and *b*'s (representing the data in their original order) can be tested for randomness on the basis of the total number of runs above and below the median, as per the procedure explained above.
(The method of runs above and below the median is helpful in testing for trends or cyclical patterns concerning economic data. In case of an upward trend, there will be first mostly *b*'s and later mostly *a*'s, but in case of a downward trend, there will be first mostly *a*'s and later mostly *b*'s. In case of a cyclical pattern, there will be a systematic alternating of *a*'s and *b*'s and probably many runs.)

Kruskal-Wallis Test:

This test is conducted in a way similar to the *U* test described above. This test is used to test the null hypothesis that '*k*' independent random samples come from identical universes against the alternative hypothesis that the means of these universes are not equal. This test is analogous to the one-way analysis of variance, but unlike the latter it does not require the assumption that the samples come from approximately normal populations or the universes having the same standard deviation.
In this test, like the *U* test, the data are ranked jointly from low to high or high to low as if they constituted a single sample. The test statistic is *H* for this test which is worked out as under:

$$H = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(n+1)$$

where $n = n_1 + n_2 + ... + n_k$ and R_i being the sum of the ranks assigned to n_i observations in the ith sample.

If the null hypothesis is true that there is no difference between the sample means and each sample has at least five items, then the sampling distribution of H can be approximated with a chi square distribution with $(k - 1)$ degrees of freedom. As such we can reject the null hypothesis at a given level of significance if H value calculated, as stated above, exceeds the concerned table value of chi-square. Let us take an example to explain the operation of this test:

Example:

Use the Kruskal-Wallis test at 5% level of significance to test the null hypothesis that a professional bowler performs equally well with the four bowling balls, given the following results:

Bowling Results In Five Games					
With Ball No A	271	282	257	248	262
With Ball No B	252	275	302	268	276
With Ball No C	260	255	239	246	266
With Ball No D	279	242	270	270	258

Solution: To apply the H test or the Kruskal-Wallis test to this problem, we begin by ranking all the given figures from the highest to the lowest, indicating besides each the name of the ball as under:

Bowling Results	Rank	Name Of The Ball Associated
302	1	B
297	2	D
282	3	A
279	4	D
276	5	B
275	6	B
271	7	A
270	8	D
268	9	B
266	10	C

262	11	A
260	12	C
258	13	D
257	14	A
255	15	C
252	16	B
248	17	A
246	18	C
242	19	D
239	20	C

For finding the values of *Ri*, we arrange the above table as under:

BALL A	RANK	BALL B	RANK	BALL C	RANK	BALL D	RANK
271	7	252	16	260	12	279	4
282	3	275	6	255	15	242	19
257	14	302	1	239	20	297	2
248	17	268	9	246	18	270	8
262	11	276	5	266	10	258	13

Now we calculate *H* statistic as under:

$$H = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(n+1)$$

$$= \frac{12}{20(20+1)} \left\{ \frac{52^2}{5} + \frac{37^2}{5} + \frac{75^2}{5} + \frac{46^2}{5} \right\} - 3(20+1)$$

$$= (0.02857)\,(2362.8) - 63 = 67.51 - 63 = 4.51$$

As the four samples have five items each, the sampling distribution of *H* approximates closely with $\chi 2$ distribution. Now taking the null hypothesis that the bowler performs equally well with the four balls, we have the value of $\chi 2 = 7.815$ for $(k - 1)$ or $4 - 1 = 3$ degrees of freedom at 5% level of significance. Since the calculated value of *H* is only 4.51 and does not exceed the $\chi 2$ value of 7.815, so we accept the null hypothesis and conclude that bowler performs equally well with the four bowling balls.