12

Testing of Hypotheses - II (Nonparametric or Distribution-free Tests)

It has already been stated in earlier chapters that a statistical test is a formal technique, based on some probability distribution, for arriving at a decision about the reasonableness of an assertion or hypothesis. The test technique makes use of one or more values obtained from sample data [often called test statistic(s)] to arrive at a probability statement about the hypothesis. But such a test technique also makes use of some more assertions about the population from which the sample is drawn. For instance, it may assume that population is normally distributed, sample drawn is a random sample and similar other assumptions. The normality of the population distribution forms the basis for making statistical inferences about the sample drawn from the population. But no such assumptions are made in case of non-parametric tests.

In a statistical test, two kinds of assertions are involved viz., an assertion directly related to the purpose of investigation and other assertions to make a probability statement. The former is an assertion to be tested and is technically called a hypothesis, whereas the set of all other assertions is called the model. When we apply a test (to test the hypothesis) without a model, it is known as distribution-free test, or the nonparametric test. Non-parametric tests do not make an assumption about the parameters of the population and thus do not make use of the parameters of the distribution. In other words, under non-parametric or distribution-free tests we do not assume that a particular distribution is applicable, or that a certain value is attached to a parameter of the population. For instance, while testing the two training methods, say *A* and *B*, for determining the superiority of one over the other, if we do not assume that the scores of the trainees are normally distributed or that the mean score of all trainees taking method *A* would be a certain value, then the testing method is known as a distribution-free or nonparametric method. In fact, there is a growing use of such tests in situations when the normality assumption is open to doubt. As a result many distribution-free tests have been developed that do not depend on the shape of the distribution or deal with the parameters of the underlying population. The present chapter discusses few such tests.

IMPORTANT NONPARAMETRIC OR DISTRIBUTION-FREE TESTS

Tests of hypotheses with 'order statistics' or 'nonparametric statistics' or 'distribution-free' statistics are known as nonparametric or distribution-free tests. The following distribution-free tests are important and generally used:

(i) Test of a hypothesis concerning some single value for the given data (such as one-sample sign test).

(ii) Test of a hypothesis concerning no difference among two or more sets of data (such as two-sample sign test, Fisher-Irwin test, Rank sum test, etc.).

(iii) Test of a hypothesis of a relationship between variables (such as Rank correlation, *Kendall's coefficient of concordance* and other tests for dependence.

(iv) Test of a hypothesis concerning variation in the given data i.e., test analogous to ANOVA viz., Kruskal-Wallis test.

(v) Tests of randomness of a sample based on the theory of runs viz., one sample runs test.

(vi) Test of hypothesis to determine if categorical data shows dependency or if two classifications are independent viz., the chi-square test. (The chi-square test has already been dealt with in Chapter 10.) The chi-square test can as well be used to make comparison between theoretical populations and actual data when categories are used.

Let us explain and illustrate some of the above stated tests which are often used in practice.

1. *Sign Tests*

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 $\mu = \mu_{H_0}$ 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_{H_0}, and with a minus (–) sign if it is less than μ_{H_0}. But if the value happens to be equal to μ_{H_0}, 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.

*If it is not possible for one reason or another to assume a symmetrical population, even then we can use the one sample sign test, but we shall then be testing the null hypothesis $\tilde{\mu} = \tilde{\mu}_{H_0}$, where $\tilde{\mu}$ is the population median.

Illustration 1

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 $\mu_{H_0} = 284$ for four rounds against the alternative hypothesis $\mu_{H_0} < 284$.

Solution: To test the null hypothesis $\mu_{H_0} = 284$ against the alternative hypothesis $\mu_{H_0} < 284$ 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:

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

$$= 0.010 + 0.001$$

(These 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:

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

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:

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

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

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

or $$0.2407$$

* Table No. 8 given in appendix at the end of the book.

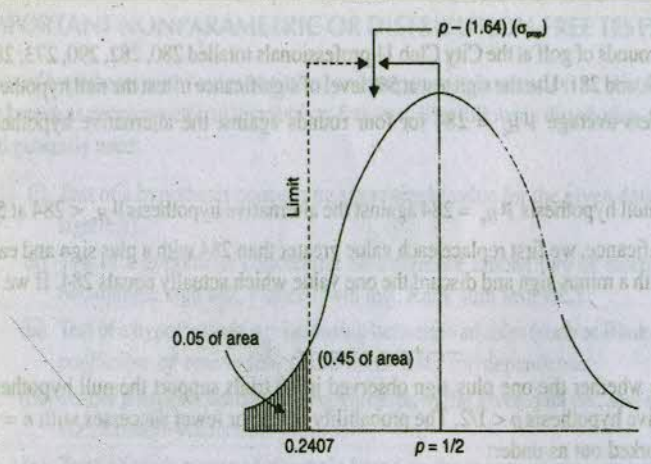

Fig. 12.1

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 of golf.

(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.

Illustration 2

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:

Table 12.1

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:

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

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:

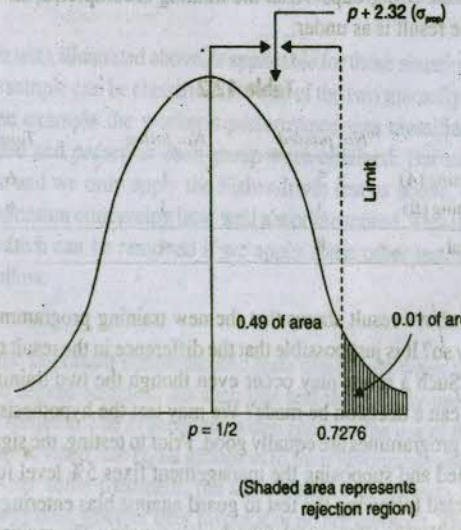

Fig. 12.2

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:

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

$$= 0.5 + 0.2276 = 0.7276$$

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 $n \cdot p$ and $n \cdot q$ both happen to be greater than 5), we can as well make use of normal approximation to binomial distribution.

2. *Fisher-Irwin Test*

Fisher-Irwin test is a distribution-free test used in testing a hypothesis concerning no difference among two sets of data. It is employed to determine whether one can reasonably assume, for example, that two supposedly different treatments are in fact different in terms of the results they produce. Suppose the management of a business unit has designed a new training programme which is now ready and as such it wishes to test its performance against that of the old training programme. For this purpose a test is performed as follows:

Twelve newly selected workers are chosen for an experiment through a standard selection procedure so that we presume that they are of equal ability prior to the experiment. This group of twelve is then divided into two groups of six each, one group for each training programme. Workers are randomly assigned to the two groups. After the training is completed, all workers are given the same examination and the result is as under:

Table 12.2

	No. passed	*No. failed*	*Total*
New Training (A)	5	1	6
Old Training (B)	3	3	6
Total	8	4	12

A casual look of the above result shows that the new training programme is superior. But the question arises: Is it really so? It is just possible that the difference in the result of the two groups may be due to chance factor. Such a result may occur even though the two training programmes were equally good. Then how can a decision be made? We may test the hypothesis for the purpose. The hypothesis is that the two programmes are equally good. Prior to testing, the significance level (or the α value) must be specified and supposing the management fixes 5% level for the purpose, which must invariably be respected following the test to guard against bias entering into the result and to avoid the possibility of vacillation oil the part of the decision maker. The required probability that the particular result or a better one for A Group would occur if the two training programmes were, in

fact, equally good, (alternatively the probability that the particular result or worse for B group would occur) be worked out. This should be done keeping in view the probability principles. For the given case, the probability that Group A has the particular result or a better one, given the null hypothesis that the two programmes are equally good, is as follows:

Pr. of Group A doing as well or better

= Pr. (5 passing and 1 failing) + Pr. (6 passing and 0 failing)

$$= \frac{{}^{8}C_5 \times {}^{4}C_1}{{}^{12}C_6} + \frac{{}^{8}C_6 \times {}^{4}C_0}{{}^{12}C_6}$$

$$= \frac{224}{924} + \frac{28}{924} = 0.24 + 0.03 = 0.27$$

Alternatively, we can work out as under:

Pr. of Group B doing as well or worse

= Pr. (3 passing and 3 failing) + Pr. (2 passing and 4 failing)

$$= \frac{{}^{8}C_3 \times {}^{4}C_3}{{}^{12}C_6} + \frac{{}^{8}C_2 \times {}^{4}C_4}{{}^{12}C_6}$$

$$= \frac{224}{924} + \frac{28}{924} = 0.24 + 0.03 = 0.27$$

Now we have to compare this calculated probability with the significance level of 5% or 0.05 already specified by the management. If we do so, we notice that the calculated value is greater than 0.05 and hence, we must accept the null hypothesis. This means that at a significance level of 5% the result obtained in the above table is not significant. Hence, we can infer that both training programmes are equally good.

This test (Fisher-Irwin test), illustrated above, is applicable for those situations where the observed result for each item in the sample can be classified into one of the two mutually exclusive categories. For instance, in the given example the worker's performance was classified as fail or pass and accordingly numbers failed and passed in each group were obtained. But supposing the score of each worker is also given and we only apply the Fisher-Irwin test as above, then certainly we are discarding the useful information concerning how well a worker scored. This in fact is the limitation of the Fisher-Irwin test which can be removed if we apply some other test, say, Wilcoxon test as stated in the pages that follow.

3. *McNemer Test*

McNemer test is one of the important nonparametric tests often used when the data happen to be nominal and relate to two related samples. As such this test is specially useful with before-after measurement of the same subjects. The experiment is designed for the use of this test in such a way that the subjects initially are divided into equal groups as to their favourable and unfavourable views about, say, any system. After some treatment, the same number of subjects are asked to express their views about the given system whether they favour it or do not favour it. Through McNemer test we in fact try to judge the significance of any observed change in views of the same subjects before

and after the treatment by setting up a table in the following form in respect of the first and second set of responses:

Table 12.3

Before treatment	*After treatment*	
	Do not favour	*Favour*
Favour	*A*	*B*
Do not favour	*C*	*D*

Since $A + D$ indicates change in people's responses ($B + C$ shows no change in responses), the expectation under null hypothesis H_0 is that $(A + D)/2$ cases change in one direction and the same proportion in other direction. The test statistic under McNemer Test is worked out as under (as it uses the under-mentioned transformation of Chi-square test):

$$\chi^2 = \frac{(|A - D| - 1)^2}{(A + D)} \text{ with d.f.} = 1$$

The minus 1 in the above equation is a correction for continuity as the Chi-square test happens to be a continuous distribution, whereas the observed data represent a discrete distribution. We illustrate this test by an example given below:

Illustration 3

In a certain before-after experiment the responses obtained from 1000 respondents, when classified, gave the following information:

Before treatment	*After treatment*	
	Unfavourable Response	*Favourable Response*
Favourable response	200 = *A*	300 = *B*
Unfavourable response	400 = *C*	100 = *D*

Test at 5% level of significance, whether there has been a significant change in people's attitude before and after the concerning experiment.

Solution: In the given question we have nominal data and the study involves before-after measurements of the two related samples, we can use appropriately the McNemer test.

We take the null hypothesis (H_0) that there has been no change in people's attitude before and after the experiment. This, in other words, means that the probability of favourable response before and unfavourable response after is equal to the probability of unfavourable response before and favourable response after i.e.,

$$H_0: P(A) = P(D)$$

We can test this hypothesis against the alternative hypothesis (H_a) viz.,

$$H_a: P(A) \neq P(D)$$

The test statistic, utilising the McNemer test, can be worked out as under:

$$\chi^2 = \frac{(|A - D| - 1)^2}{(A + D)} = \frac{(|200 - 100| - 1)^2}{(200 + 100)}$$

$$= \frac{99 \times 99}{300} = 32.67$$

Degree of freedom = 1.

From the Chi-square distribution table, the value of χ^2 for 1 degree of freedom at 5% level of significance is 3.84. The calculated value of χ^2 is 32.67 which is greater than the table value, indicating that we should reject the null hypothesis. As such we conclude that the change in people's attitude before and after the experiment is significant.

4. *Wilcoxon Matched-pairs Test (or Signed Rank Test)*

In various research situations in the context of two-related samples (i.e., case of matched paires such as a study where husband and wife are matched or when we compare the output of two similar machines or where some subjects are studied in context of before-after experiment) when we can determine both direction and magnitude of difference between matched values, we can use an important non-parametric test viz., Wilcoxon matched-paires test. While applying this test, we first find the differences (d_i) between each pair of values and assign rank to the differences from the smallest to the largest without regard to sign. The actual signs of each difference are then put to corresponding ranks and the test statistic T is calculated which happens to be the smaller of the two sums viz., the sum of the negative ranks and the sum of the positive ranks.

While using this test, we may come across two types of tie situations. One situation arises when the two values of some matched pair(s) are equal i.e., the difference between values is zero in which case we drop out the pair(s) from our calculations. The other situation arises when two or more pairs have the same difference value in which case we assign ranks to such pairs by averaging their rank positions. For instance, if two pairs have rank score of 5, we assign the rank of 5.5 i.e., $(5 + 6)/2 = 5.5$ to each pair and rank the next largest difference as 7.

When the given number of matched pairs after considering the number of dropped out pair(s), if any, as stated above is equal to or less than 25, we use the table of critical values of T (Table No. 7 given in appendix at the end of the book) for the purpose of accepting or rejecting the null hypothesis of no difference between the values of the given pairs of observations at a desired level of significance. For this test, the calculated value of T must be equal to or smaller than the table value in order to reject the null hypothesis. In case the number exceeds 25, the sampling distribution of T is taken as approximately normal with mean $U_T = n(n + 1)/4$ and standard deviation

$$\sigma_T = \sqrt{n(n + 1)(2n + 1)/24},$$

where n = [(number of given matched pairs) – (number of dropped out pairs, if any)] and in such situation the test statistic z is worked out as under:

$$z = \frac{T - U_T}{\sigma_T}$$

We may now explain the use of this test by an example.

Illustration 4

An experiment is conducted to judge the effect of brand name on quality perception. 16 subjects are recruited for the purpose and are asked to taste and compare two samples of product on a set of scale items judged to be ordinal. The following data are obtained:

Pair	Brand A	Brand B
1	73	51
2	43	41
3	47	43
4	53	41
5	58	47
6	47	32
7	52	24
8	58	58
9	38	43
10	61	53
11	56	52
12	56	57
13	34	44
14	55	57
15	65	40
16	75	68

Test the hypothesis, using Wilcoxon matched-pairs test, that there is no difference between the perceived quality of the two samples. Use 5% level of significance.

Solution: Let us first write the null and alternative hypotheses as under:

H_0: There is no difference between the perceived quality of two samples.

H_a: There is difference between the perceived quality of the two samples.

Using Wilcoxon matched-pairs test, we work out the value of the test statistic T as under:

Table 12.4

Pair	Brand A	Brand B	Difference d_i	Rank of $\lvert d_i \rvert$	Rank with signs +	Rank with signs −
1	73	51	22	13	13	...
2	43	41	2	2.5	2.5	...

Contd.

Pair	Brand A	Brand B	Difference d_i	Rank of $\lvert d_i \rvert$	Rank with signs +	Rank with signs −
3	47	43	4	4.5	4.5	...
4	53	41	12	11	11	...
5	58	47	11	10	10	...
6	47	32	15	12	12	...
7	52	24	28	15	15	...
8	58	58	0	–	–	–
9	38	43	−5	6	...	−6
10	61	53	8	8	8	...
11	56	52	4	4.5	4.5	...
12	56	57	−1	1	...	−1
13	34	44	−10	9	...	−9
14	55	57	−2	2.5	...	−2.5
15	65	40	25	14	14	...
16	75	68	7	7	7	...
				TOTAL	101.5	−18.5
				Hence,	$T = 18.5$	

We drop out pair 8 as 'd' value for this is zero and as such our $n = (16 - 1) = 15$ in the given problem.

The table value of T at five percent level of significance when $n = 15$ is 25 (using a two-tailed test because our alternative hypothesis is that there is difference between the perceived quality of the two samples). The calculated value of T is 18.5 which is less than the table value of 25. As such we reject the null hypothesis and conclude that there is difference between the perceived quality of the two samples.

5. *Rank Sum Tests*

Rank sum tests are a whole family of test, but we shall describe only two such tests commonly used viz., the U test and the H test. U test is popularly known as Wilcoxon-Mann-Whitney test, whereas H test is also known as Kruskal-Wallis test. A brief description of the said two tests is given below:

(a) *Wilcoxon-Mann-Whitney test (or U-test):* This is a very popular test amongst the rank sum tests. This test is used to determine whether two independent samples have been drawn from the same population. It uses more information than the sign test or the Fisher-Irwin test. This test applies under very general conditions and requires only that the populations sampled are continuous. However, in practice even the violation of this assumption does not affect the results very much.

To perform this test, we first of all rank the data jointly, taking them as belonging to a single sample in either an increasing or decreasing order of magnitude. We usually adopt low to high ranking process which means we assign rank 1 to an item with lowest value, rank 2 to the next higher item and so on. In case there are ties, then we would assign each of the tied observation the mean of the ranks which they jointly occupy. For example, if sixth, seventh and eighth values are identical, we would assign each the rank $(6 + 7 + 8)/3 = 7$. After this we find the sum of the ranks assigned to the

values of the first sample (and call it R_1) and also the sum of the ranks assigned to the values of the second sample (and call it R_2). Then we work out the test statistic i.e., U, which is a measurement of the difference between the ranked observations of the two samples as under:

$$U = n_1 \cdot n_2 + \frac{n_1(n_1 + 1)}{2} - R_1$$

where n_1, and n_2 are the sample sizes and R_1 is the sum of ranks assigned to the values of the first sample. (In practice, whichever rank sum can be conveniently obtained can be taken as R_1, since it is immaterial which sample is called the first sample.)

In applying U-test we take the null hypothesis that the two samples come from identical populations. If this hypothesis is true, it seems reasonable to suppose that the means of the ranks assigned to the values of the two samples should be more or less the same. Under the alternative hypothesis, the means of the two populations are not equal and if this is so, then most of the smaller ranks will go to the values of one sample while most of the higher ranks will go to those of the other sample.

If the null hypothesis that the $n_1 + n_2$ observations came from identical populations is true, the said 'U' statistic has a sampling distribution with

$$\text{Mean} = \mu_U = \frac{n_1 \cdot n_2}{2}$$

and Standard deviation (or the standard error)

$$= \sigma_U = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}$$

If n_1 and n_2 are sufficiently large (i.e., both greater than 8), the sampling distribution of U can be approximated closely with normal distribution and the limits of the acceptance region can be determined in the usual way at a given level of significance. But if either n_1 or n_2 is so small that the normal curve approximation to the sampling distribution of U cannot be used, then exact tests may be based on special tables such as one given in the, appendix,* showing selected values of Wilcoxon's (unpaired) distribution. We now can take an example to explain the operation of U test.

Illustration 5

The values in one sample are 53, 38, 69, 57, 46, 39, 73, 48, 73, 74, 60 and 78. In another sample they are 44, 40, 61, 52, 32, 44, 70, 41, 67, 72, 53 and 72. Test at the 10% level the hypothesis that they come from populations with the same mean. Apply U-test.

Solution: First of all we assign ranks to all observations, adopting low to high ranking process on the presumption that all given items belong to a single sample. By doing so we get the following:

*Table No. 6 given in appendix at the end of the book.

Table 12.5

Size of sample item in ascending order	Rank	Name of related sample: [A for sample one and B for sample two]
32	1	B
38	2	A
39	3	A
40	4	B
41	5	B
44	6.5	B
44	6.5	B
46	8	A
48	9	A
52	10	B
53	11.5	B
53	11.5	A
57	13	A
60	14	A
61	15	B
67	16	B
69	17	A
70	18	B
72	19.5	B
72	19.5	B
73	21.5	A
73	21.5	A
74	23	A
78	24	A

From the above we find that the sum of the ranks assigned to sample one items or $R_1 = 2 + 3 + 8 + 9 + 11.5 + 13 + 14 + 17 + 21.5 + 21.5 + 23 + 24 = 167.5$ and similarly we find that the sum of ranks assigned to sample two items or $R_2 = 1 + 4 + 5 + 6.5 + 6.5 + 10 + 11.5 + 15 + 16 + 18 + 19.5 + 19.5 = 132.5$ and we have $n_1 = 12$ and $n_2 = 12$

Hence, test statistic $U = n_1 \cdot n_2 + \frac{n_1(n_1 + 1)}{2} - R_1$

$$= (12)(12) + \frac{12(12 + 1)}{2} - 167.5$$

$$= 144 + 78 - 167.5 = 54.5$$

Since in the given problem n_1 and n_2 both are greater than 8, so the sampling distribution of U approximates closely with normal curve. Keeping this in view, we work out the mean and standard deviation taking the null hypothesis that the two samples come from identical populations as under:

$$\mu_U = \frac{n_1 \times n_2}{2} = \frac{(12)(12)}{2} = 72$$

$$\sigma_U = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}} = \sqrt{\frac{(12)(12)(12 + 12 + 1)}{12}}$$

$$= 17.32$$

As the alternative hypothesis is that the means of the two populations are not equal, a two-tailed test is appropriate. Accordingly the limits of acceptance region, keeping in view 10% level of significance as given, can be worked out as under:

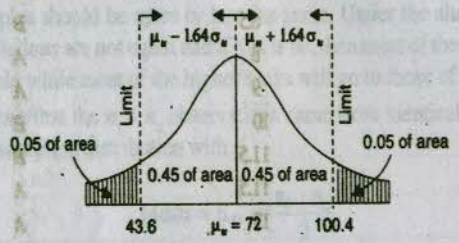

Fig. 12.3

As the z value for 0.45 of the area under the normal curve is 1.64, we have the following limits of acceptance region:

Upper limit $= \mu_U + 1.64\,\sigma_U = 72 + 1.64(17.32) = 100.40$

Lower limit $= \mu_U - 1.64\,\sigma_U = 72 - 1.64(17.32) = 43.60$

As the observed value of U is 54.5 which is in the acceptance region, we accept the null hypothesis and conclude that the two samples come from identical populations (or that the two populations have the same mean) at 10% level.

We can as well calculate the U statistic as under using R_2 value:

$$U = n_1 \cdot n_2 + \frac{n_2(n_2 + 1)}{2} - R_2$$

$$= (12)(12) + \frac{12(12 + 1)}{2} - 132.5$$

$$= 144 + 78 - 132.5 = 89.5$$

The value of U also lies in the acceptance region and as such our conclusion remains the same, even if we adopt this alternative way of finding U.

We can take one more example concerning U test wherein n_1 and n_2 are both less than 8 and as such we see the use of table given in the appendix concerning values of Wilcoxon's distribution (unpaired distribution).

Illustration 6

Two samples with values 90, 94, 36 and 44 in one case and the other with values 53, 39, 6, 24, and 33 are given. Test applying Wilcoxon test whether the two samples come from populations with the same mean at 10% level against the alternative hypothesis that these samples come from populations with different means.

Solution: Let us first assign ranks as stated earlier and we get:

Table 12.6

Size of sample item in ascending order	*Rank*	*Name of related sample (Sample one as A Sample two as B)*
6	1	B
24	2	B
33	3	B
36	4	A
39	5	B
44	6	A
53	7	B
90	8	A
94	9	A

Sum of ranks assigned to items of sample one = 4 + 6 + 8 + 9 = 27

No. of items in this sample = 4

Sum of ranks assigned to items of sample two = 1 + 2 + 3 + 5 + 7 = 18

No. of items in this sample = 5

As the number of items in the two samples is less than 8, we cannot use the normal curve approximation technique as stated above and shall use the table giving values of Wilcoxon's distribution. To use this table, we denote 'W_s' as the smaller of the two sums and 'W_l' the larger. Also, let 's' be the number of items in the sample with smaller sum and let 'l' be the number of items in the sample with the larger sum. Taking these notations we have for our question the following values:

$$W_s = 18;\ s = 5;\ W_l = 27;\ l = 4$$

The value of W_s is 18 for sample two which has five items and as such $s = 5$. We now find the difference between W_s and the minimum value it might have taken, given the value of s. The minimum value that W_s could have taken, given that $s = 5$, is the sum of ranks 1 through 5 and this comes as equal to 1 + 2 + 3 + 4 + 5 = 15. Thus, (W_s – Minimum W_s) = 18 – 15 = 3. To determine the probability that a result as extreme as this or more so would occur, we find the cell of the table which is in the column headed by the number 3 and in the row for $s = 5$ and $l = 4$ (the specified values of l are given in the second column of the table). The entry in this cell is 0.056 which is the required probability of getting a value as small as or smaller than 3 and now we should compare it with the significance level of 10%. Since the alternative hypothesis is that the two samples come from populations with different means, a two-tailed test is appropriate and accordingly 10% significance level will mean 5% in the left tail and 5% in the right tail. In other words, we should compare the calculated probability with the

probability of 0.05, given the null hypothesis and the significance level. If the calculated probability happens to be greater than 0.05 (which actually is so in the given case as 0.056 > 0.05), then we should accept the null hypothesis. Hence, in the given problem, we must conclude that the two samples come from populations with the same mean.

(The same result we can get by using the value of W_l. The only difference is that the value maximum $W_l - W_l$ is required. Since for this problem, the maximum value of W_l (given $s = 5$ and $l = 4$) is the sum of 6 through 9 i.e., $6 + 7 + 8 + 9 = 30$, we have Max. $W_l - W_l = 30 - 27 = 3$ which is the same value that we worked out earlier as W_s – Minimum W_s. All other things then remain the same as we have stated above).

(b) *The Kruskal-Wallis test (or H 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 $n = n_1 + n_2 + \ldots + n_k$ and R_i being the sum of the ranks assigned to n_i observations in the *i*th 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:

Illustration 7

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	297	270	258

* If any of the given samples has less than five items then chi-square distribution approximation can not be used and the exact tests may be based on table meant for it given in the book "Non-parametric statistics for the behavioural sciences" by S. Siegel.

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:

Table 12.7

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 R_i, we arrange the above table as under:

Table 12.7 (a): Bowling Results with Different Balls and Corresponding Rank

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	158	13
$n_1 = 5$	$R_1 = 52$	$n_2 = 5$	$R_2 = 37$	$n_3 = 5$	$R_3 = 75$	$n_4 = 5$	$R_4 = 46$

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 χ^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 χ^2 value of 7.815, so we accept the null hypothesis and conclude that bowler performs equally well with the four bowling balls.

6. *One Sample 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)

*For the application of H test, it is not necessary that all samples should have equal number of items.

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 $\sigma_r = \sqrt{2n_1n_2 \dfrac{2n_1n_2 - n_1 - n_2}{(n_1+n_2)^2\,(n_1+n_2-1)}}$

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.

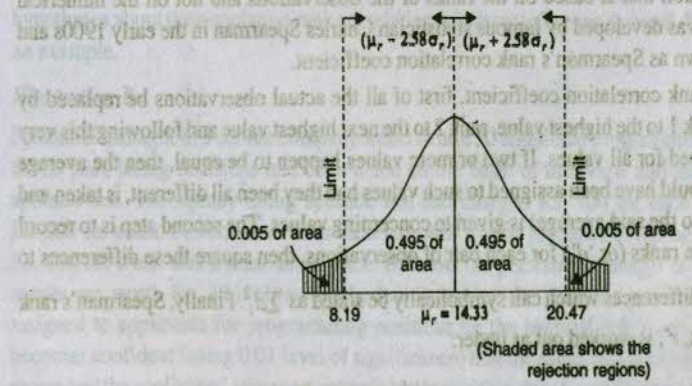

Fig. 12.4

*This assumption can be applied when n_1 and n_2 are sufficiently large i.e., they should not be less than 10. But in case n_1 or n_2 is so small that the normal curve approximation assumption cannot be used, then exact tests may be based on special tables which can be seen in the book *Non-parametric Statistics for the Behavioural Science* by S. Siegel.

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 = $\mu_r + (2.58)(2.38) = 14.33 + 6.14 = 20.47$ and

Lower limit = $\mu_r - (2.58)(2.38) = 14.33 - 6.14 = 8.19$

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., $\alpha = 0.01$. 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.)

7. Spearman's Rank Correlation

When the data are not available to use in numerical form for doing correlation analysis but when the information is sufficient to rank the data as first, second, third, and so forth, we quite often use the rank correlation method and work out the coefficient of rank correlation. In fact, the rank correlation coefficient is a measure of correlation that exists between the two sets of ranks. In other words, it is a measure of association that is based on the ranks of the observations and not on the numerical values of the data. It was developed by famous statistician Charles Spearman in the early 1900s and as such it is also known as Spearman's rank correlation coefficient.

For calculating rank correlation coefficient, first of all the actual observations be replaced by their ranks, giving rank 1 to the highest value, rank 2 to the next highest value and following this very order ranks are assigned for all values. If two or more values happen to be equal, then the average of the ranks which should have been assigned to such values had they been all different, is taken and the same rank (equal to the said average) is given to concerning values. The second step is to record the difference between ranks (or '*d*') for each pair of observations, then square these differences to obtain a total of such differences which can symbolically be stated as $\sum d_i^2$. Finally, Spearman's rank correlation coefficient, r^*, is worked out as under:

$$\text{Spearman's } r = 1 - \left\{\frac{6\sum d_i^2}{n(n^2-1)}\right\}$$

*Some authors use the symbol Rho (ρ) for this coefficient. Rho is to be used when the sample size does not exceed 30.

where n = number of paired observations.

The value of Spearman's rank correlation coefficient will always vary between ± 1, $+1$, indicating a perfect positive correlation and -1 indicating perfect negative correlation between two variables. All other values of correlation coefficient will show different degrees of correlation.

Suppose we get $r = 0.756$ which suggests a substantial positive relationship between the concerning two variables. But how we should test this value of 0.756? The testing device depends upon the value of n. For small values of n (i.e., n less than 30), the distribution of r is not normal and as such we use the table showing the values for Spearman's Rank correlation (Table No. 5 given in Appendix at the end of the book) to determine the acceptance and rejection regions. Suppose we get $r = 0.756$ for a problem where $n = 15$ and want to test at 5% level of significance the null hypothesis that there is zero correlation in the concerning ranked data. In this case our problem is reduced to test the null hypothesis that there is no correlation i.e., $\mu_r = 0$ against the alternative hypothesis that there is a correlation i.e., $\mu_r \neq 0$ at 5% level. In this case a two-tailed test is appropriate and we look in the said table in row for $n = 15$ and the column for a significance level of 0.05 and find that the critical values for r are ± 0.5179 i.e., the upper limit of the acceptance region is 0.5179 and the lower limit of the acceptance region is -0.5179. And since our calculated $r = 0.756$ is outside the limits of the acceptance region, we reject the null hypothesis and accept the alternative hypothesis that there is a correlation in the ranked data.

In case the sample consists of more than 30 items, then the sampling distribution of r is approximately normal with a mean of zero and a standard deviation of $1/\sqrt{n-1}$ and thus, the standard error of r is:

$$\sigma_r = \frac{1}{\sqrt{n-1}}$$

We can use the table of area under normal curve to find the appropriate z values for testing hypotheses about the population rank correlation and draw inference as usual. We can illustrate it, by an example.

Illustration 8

Personnel manager of a certain company wants to hire 30 additional programmers for his corporation. In the past, hiring decisions had been made on the basis of interview and also on the basis of an aptitude test. The agency doing aptitude test had charged Rs. 100 for each test, but now wants Rs. 200 for a test. Performance on the test has been a good predictor of a programmer's ability and Rs. 100 for a test was a reasonable price. But now the personnel manager is not sure that the test results are worth Rs. 200. However, he has kept over the past few years records of the scores assigned to applicants for programming positions on the basis of interviews taken by him. If he becomes confident (using 0.01 level of significance) that the rank correlation between his interview scores and the applicants' scores on aptitude test is positive, then he will feel justified in discontinuing the aptitude test in view of the increased cost of the test. What decision should he take on the basis of the following sample data concerning 35 applicants?

Sample Data Concerning 35 Applicants

Serial Number	Interview score	Aptitude test score
1	81	113
2	88	88
3	55	76
4	83	129
5	78	99
6	93	142
7	65	93
8	87	136
9	95	82
10	76	91
11	60	83
12	85	96
13	93	126
14	66	108
15	90	95
16	69	65
17	87	96
18	68	101
19	81	111
20	84	121
21	82	83
22	90	79
23	63	71
24	78	109
25	73	68
26	79	121
27	72	109
28	95	121
29	81	140
30	87	132
31	93	135
32	85	143
33	91	118
34	94	147
35	94	138

Solution: To solve this problem we should first work out the value of Spearman's *r* as under:

Table 12.8: Calculation of Spearman's

S. No.	Interview score X	Aptitude test score Y	Rank X	Rank Y	Rank Difference 'd_i' (Rank X) – (Rank Y)	Differences squared d_i^2
1	81	113	21	15	6	36
2	88	88	11	27	–16	256
3	55	76	35	32	3	9
4	83	129	18	9	9	81
5	78	99	24.5	21	3.5	12.25
6	93	142	6	3	3	9
7	65	93	32	25	7	49
8	87	136	13	6	7	49
9	95	82	1.5	30	–28.5	812.25
10	76	91	26	26	0	0
11	60	83	34	28.5	5.5	30.25
12	85	96	15.5	22.5	–7	49
13	93	126	6	10	–4	16
14	66	108	31	18.5	12.5	156.25
15	90	95	9.5	24	–14.5	210.25
16	69	65	29	35	–6	36
17	87	96	13	22.5	–9.5	90.25
18	68	101	30	20	10	100
19	81	111	21	16	5	25
20	84	121	17	12	5	25
21	82	83	19	28.5	–9.5	90.25
22	90	79	9.5	31	–21.5	462.25
23	63	71	33	33	0	0
24	78	108	24.5	18.5	6	36
25	73	68	27	34	–7	49
26	79	121	23	12	11	121
27	72	109	28	17	11	121
28	95	121	1.5	12	–10.5	110.25
29	81	140	21	4	17	289
30	87	132	13	8	5	25
31	93	135	6	7	–1	1
32	85	143	15.5	2	13.5	182.25
33	91	118	8	14	–6	36
34	94	147	3.5	1	2.5	6.25
35	94	138	3.5	5	–1.5	225
$n = 35$						$\Sigma d_i^2 = 3583$

$$\text{Spearman's } r = 1 - \left\{\frac{6\sum d_i^2}{n(n^2-1)}\right\} = 1 - \left\{\frac{6 \times 3583}{35(35^2-1)}\right\}$$

$$= 1 - \frac{21498}{42840} = 0.498$$

Since $n = 35$ the sampling distribution of r is approximately normal with a mean of zero and a standard deviation of $1/\sqrt{n-1}$. Hence the standard error of r is

$$\sigma_r = \frac{1}{\sqrt{n-1}} = \frac{1}{\sqrt{35-1}} = 0.1715$$

As the personnel manager wishes to test his hypothesis at 0.01 level of significance, the problem can be stated:

Null hypothesis that there is no correlation between interview score and aptitude test score i.e., $\mu_r = 0$.

Alternative hypothesis that there is positive correlation between interview score and aptitude test score i.e., $\mu_r > 0$.

As such one-tailed test is appropriate which can be indicated as under in the given case:

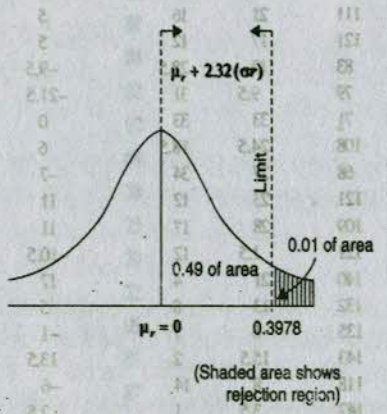

Fig. 12.5

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 alternative hypothesis is of > type) of the acceptance region as under:

$$\mu_r + (2.32)(0.1715)$$
$$= 0 + 0.3978$$
$$= 0.3978$$

We now find the observed $r = 0.498$ and as such it comes in the rejection region and, therefore, we reject the null hypothesis at 1% level and accept the alternative hypothesis. Hence we conclude that correlation between interview score and aptitude test score is positive. Accordingly personnel manager should decide that the aptitude test be discontinued.

8. Kendall's Coefficient of Concordance

Kendall's coefficient of concordance, represented by the symbol W, is an important non-parametric measure of relationship. It is used for determining the degree of association among several (k) sets of ranking of N objects or individuals. When there are only two sets of rankings of N objects, we generally work out Spearman's coefficient of correlation, but Kendall's coefficient of concordance (W) is considered an appropriate measure of studying the degree of association among three or more sets of rankings. This descriptive measure of the agreement has special applications in providing a standard method of ordering objects according to consensus when we do not have an objective order of the objects.

The basis of Kendall's coefficient of concordance is to imagine how the given data would look if there were no agreement among the several sets of rankings, and then to imagine how it would look if there were perfect agreement among the several sets. For instance, in case of, say, four interviewers interviewing, say, six job applicants and assigning rank order on suitability for employment, if there is observed perfect agreement amongst the interviewers, then one applicant would be assigned rank 1 by all the four and sum of his ranks would be $1 + 1 + 1 + 1 = 4$. Another applicant would be assigned a rank 2 by all four and the sum of his ranks will be $2 + 2 + 2 + 2 = 8$. The sum of ranks for the six applicants would be 4, 8, 12, 16, 20 and 24 (not necessarily in this very order). In general, when perfect agreement exists among ranks assigned by k judges to N objects, the rank sums are k, $2k$, $3k$, ... Nk. The, total sum of N ranks for k judges is $kN(N+1)/2$ and the mean rank sum is $k(N+1)/2$. The degree of agreement between judges reflects itself in the variation in the rank sums. When all judges agree, this sum is a maximum. Disagreement between judges reflects itself in a reduction in the variation of rank sums. For maximum disagreement, the rank sums will tend to be more or less equal. This provides the basis for the definition of a coefficient of concordance. When perfect agreement exists between judges, W equals to 1. When maximum disagreement exists, W equals to 0. It may be noted that W does not take negative values because of the fact that with more than two judges complete disagreement cannot take place. Thus, coefficient of concordance (W) is an index of divergence of the actual agreement shown in the data from the perfect agreement.

The procedure for computing and interpreting Kendall's coefficient of concordance (W) is as follows:

(a) All the objects, N, should be ranked by all k judges in the usual fashion and this information may be put in the form of a k by N matrix;

(b) For each object determine the sum of ranks (R_j) assigned by all the k judges;

(c) Determine $\bar{R}_j$ and then obtain the value of s as under:

$$s = \Sigma\left(R_j - \bar{R}_j\right)^2$$

(d) Work out the value of W using the following formula:

$$W = \frac{s}{\frac{1}{12}k^2\left(N^3 - N\right)}$$

where $s = \Sigma\left(R_j - \bar{R}_j\right)^2$;

k = no. of sets of rankings i.e., the number of judges;

N = number of objects ranked;

$\frac{1}{12}k^2\left(N^3 - N\right)$ = maximum possible sum of the squared deviations i.e., the sum s which would occur with perfect agreement among k rankings.

Case of Tied Ranks

Where tied ranks occur, the average method of assigning ranks be adopted i.e., assign to each member the average rank which the tied observations occupy. If the ties are not numerous, we may compute 'W' as stated above without making any adjustment in the formula; but if the ties are numerous, a correction factor is calculated for each set of ranks. This correction fact is

$$T = \frac{\Sigma\left(t^3 - t\right)}{12}$$

where t = number of observations in a group tied for a given rank.

For instance, if the ranks on X are 1, 2, 3.5, 5, 6, 3.5, 8, 10, 8, 8, we have two groups of ties, one of two ranks and one of three ranks. The correction factor for this set of ranks for X would be

$$T = \frac{\left(2^3 - 2\right) + \left(3^3 - 3\right)}{12} = 2.5$$

A correction factor T is calculated for each of the k sets of ranks and these are added together over the k sets to obtain ΣT. We then use the formula for finding the value of 'W' as under:

$$W = \frac{s}{\frac{1}{12}k^2\left(N^3 - N\right) - k\Sigma T}$$

The application of the correction in this formula tends to increase the size of W, but the correction factor has a very limited effect unless the ties are quite numerous.

(e) The method for judging whether the calculated value of W is significantly different from zero depends on the size of N as stated below:

(i) If N is 7 or smaller, Table No. 9 given in appendix at the end of the book gives critical values of s associated with W's significance at 5% and 1% levels. If an observed s is equal to or greater than that shown in the table for a particular level of significance, then H_0 T (i.e., k sets of rankings are independent) may be rejected at that level of significance.

(ii) If N is larger than 7, we may use χ^2 value to be worked out as: $\chi^2 = k(N-1)$. W with d.f. $= (N-1)$ for judging W's significance at a given level in the usual way of using χ^2 values.

(f) Significant value of W may be interpreted and understood as if the judges are applying essentially the same standard in ranking the N objects under consideration, but this should never mean that the orderings observed are correct for the simple reason that all judges can agree in ordering objects because they all might employ 'wrong' criterion. Kendall, therefore, suggests that the best estimate of the 'true' rankings of N objects is provided, when W is significant, by the order of the various sums of ranks, R_j. If one accepts the criterion which the various judges have agreed upon, then the best estimate of the 'true' ranking is provided by the order of the sums of ranks. The best estimate is related to the lowest value observed amongst R_j.

This can be illustrated with the help of an example.

Illustration 9

Seven individuals have been assigned ranks by four judges at a certain music competition as shown in the following matrix:

	Individuals						
	A	*B*	*C*	*D*	*E*	*F*	*G*
Judge 1	1	3	2	5	7	4	6
Judge 2	2	4	1	3	7	5	6
Judge 3	3	4	1	2	7	6	5
Judge 4	1	2	5	4	6	3	7

Is there significant agreement in ranking assigned by different judges? Test at 5% level. Also point out the best estimate of the true rankings.

Solution: As there are four sets of rankings, we can work out the coefficient of concordance (W) for judging significant agreement in ranking by different judges. For this purpose we first develop the given matrix as under:

Table 12.9

$K = 4$	*Individuals*							$\therefore N = 7$
	A	*B*	*C*	*D*	*E*	*F*	*G*	
Judge 1	1	3	2	5	7	4	6	
Judge 2	2	4	1	3	7	5	6	
Judge 3	3	4	1	2	7	6	5	
Judge 4	1	2	5	4	6	3	7	
Sum of ranks (R_j)	7	13	9	14	27	18	24	$\Sigma R_j = 112$
$\left(R_j - \bar{R}_j\right)^2$	81	9	49	4	121	4	64	$\therefore s = 332$

$$\because \quad \bar{R}_j = \frac{\Sigma R_j}{N} = \frac{112}{7} = 16$$

$$\therefore \quad s = 332$$

$$\therefore \quad W = \frac{s}{\frac{1}{12}k^2\left(N^3 - N\right)} = \frac{332}{\frac{1}{12}(4)^2\left(7^3 - 7\right)} = \frac{332}{\frac{16}{12}(336)} = \frac{332}{448} = 0.741$$

To judge the significance of this W, we look into the Table No. 9 given in appendix for finding the value of s at 5% level for $k = 4$ and $N = 7$. This value is 217.0 and thus for accepting the null hypothesis (H_0) that k sets of rankings are independent) our calculated value of s should be less than 217. But the worked out value of s is 332 which is higher than the table value which fact shows that $W = 0.741$ is significant. Hence, we reject the null hypothesis and infer that the judges are applying essentially the same standard in ranking the N objects i.e., there is significant agreement in ranking by different judges at 5% level in the given case. The lowest value observed amongst R_j is 7 and as such the best estimate of true rankings is in the case of individual A i.e., all judges on the whole place the individual A as first in the said music competition.

Illustration 10

Given is the following information:

$k = 13$

$N = 20$

$W = 0.577$

Determine the significance of W at 5% level.

Solution: As N is larger than 7, we shall workout the value of χ^2 for determining W's significance as under:

$$\chi^2 = k(N-1)W \text{ with } N-1 \text{ degrees of freedom}$$

$$\therefore \quad \chi^2 = 13(20-1)(0.577)$$

or $$\chi^2 = (247)(0.577) = 142.52$$

Table value of χ^2 at 5% level for $N - 1 = 20 - 1 = 19$ d.f. is 30.144 but the calculated value of χ^2 is 142.52 and this is considerably higher than the table value. This does not support the null hypothesis of independence and as such we can infer that W is significant at 5% level.

RELATIONSHIP BETWEEN SPEARMAN'S *r*'s AND KENDALL'S *W*

As stated above, W is an appropriate measure of studying the degree of association among three or more sets of ranks, but we can as well determine the degree of association among k sets of rankings by averaging the Spearman's correlation coefficients (r's) between all possible pairs (i.e., ${}^kC_2$ or $k(k-1)/2$) of rankings keeping in view that W bears a linear relation to the average r's taken over

all possible pairs. The relationship between the average of Spearman's r's and Kendall's W can be put in the following form:

$$\text{average of } r\text{'s} = (kW - 1)/(k - 1)$$

But the method of finding W using average of Spearman's r's between all possible pairs is quite tedious, particularly when k happens to be a big figure and as such this method is rarely used in practice for finding W.

Illustration 11

Using data of illustration No. 9 above, find W using average of Spearman's r's.

Solution: As $k = 4$ in the given question, the possible pairs are equal to $k(k-1)/2 = 4(4-1)/2 = 6$ and we work out Spearman's r for each of these pairs as shown in Table 12.10.

Now we can find W using the following relationship formula between r's average and W

$$\text{Average of } r\text{'s} = (kW - 1)/(k - 1)$$

or $$0.655 = (4W - 1)/(4 - 1)$$

or $$(0.655)(3) = 4W - 1$$

or $$W = \frac{(0.655)(3) + 1}{4} = \frac{2.965}{4} = 0.741$$

[*Note:* This value of W is exactly the same as we had worked out using the formula:

$$W = s/[(1/12)(k^2)(N^3 - N)]$$

CHARACTERISTICS OF DISTRIBUTION-FREE OR NON-PARAMETRIC TESTS

From what has been stated above in respect of important non-parametric tests, we can say that these tests share in main the following characteristics:

1. They do not suppose any particular distribution and the consequential assumptions.
2. They are rather quick and easy to use i.e., they do not require laborious computations since in many cases the observations are replaced by their rank order and in many others we simply use signs.
3. They are often not as efficient or 'sharp' as tests of significance or the parametric tests. An interval estimate with 95% confidence may be twice as large with the use of non-parametric tests as with regular standard methods. The reason being that these tests do not use all the available information but rather use groupings or rankings and the price we pay is a loss in efficiency. In fact, when we use non-parametric tests, we make a trade-off: we loose sharpness in estimating intervals, but we gain the ability to use less information and to calculate faster.
4. When our measurements are not as accurate as is necessary for standard tests of significance, then non-parametric methods come to our rescue which can be used fairly satisfactorily.
5. Parametric tests cannot apply to ordinal or nominal scale data but non-parametric tests do not suffer from any such limitation.
6. The parametric tests of difference like 't' or 'F' make assumption about the homogeneity of the variances whereas this is not necessary for non-parametric tests of difference.

Table 12.10: Difference between Ranks $|d|$ Assigned by $k = 4$ Judges and the Square Values of such Differences (d_i^2) for all Possible Pairs of Judges

Individuals	Pair 1 – 2		Pair 1 – 3		Pair 1 – 4		Pair 2 – 3		Pair 2 – 4		Pair 3 – 4	
	$\|d\|$	d^2	$\|d\|$	d^2	$\|d\|$	d^2	$\|d\|$	d^2	$\|d\|$	d^2	$\|d\|$	d^2
A	1	1	2	4	0	0	1	1	1	1	2	4
B	1	1	1	1	1	1	0	0	2	4	2	4
C	–1	1	1	1	3	9	0	0	4	16	4	16
D	–2	4	3	9	1	1	1	1	1	1	2	4
E	0	0	0	0	1	1	0	0	1	1	1	1
F	1	1	2	4	1	1	1	1	2	4	3	9
G	0	0	1	1	1	1	1	1	1	1	2	4
		$\Sigma d_i^2 = 8$		$\Sigma d_i^2 = 20$		$\Sigma d_i^2 = 14$		$\Sigma d_i^2 = 4$		$\Sigma d_i^2 = 28$		$\Sigma d_i^2 = 42$
Spearman's Coefficient of Correlation												
$r = 1 - \frac{6\Sigma d_i^2}{N(N^2-1)}$		$r_{12} = 0.857$		$r_{13} = 0.643$		$r_{14} = 0.750$		$r_{23} = 0.929$		$r_{24} = 0.500$		$r_{34} = 0.250$

$$\text{Average of Spearman's } r\text{'s} = \frac{0.857 + 0.643 + 0.750 + 0.929 + 0.500 + 0.250}{6}$$

$$= \frac{3.929}{6} = 0.655$$

CONCLUSION

There are many situations in which the various assumptions required for standard tests of significance (such as that population is normal, samples are independent, standard deviation is known, etc.) cannot be met, then we can use non-parametric methods. Moreover, they are easier to explain and easier to understand. This is the reason why such tests have become popular. But one should not forget the fact that they are usually less efficient/powerful as they are based on no assumption (or virtually no assumption) and we all know that the less one assumes, the less one can infer from a set of data. But then the other side must also be kept in view that the more one assumes, the more one limits the applicability of one's methods.

Questions

1. Give your understanding of non-parametric or distribution free methods explaining their important characteristics.
2. Narrate the various advantages of using non-parametric tests. Also point out their limitations.
3. Briefly describe the different non-parametric tests explaining the significance of each such test.
4. On 15 occasions Mr. Kalicharan had to wait 4, 8, 2, 7, 7, 5, 8, 6, 1, 9, 6, 6, 5, 9 and 5 minutes for the bus he takes to reach his office. Use the sign test at 5% level of significance to test the bus company's claim that on the average Mr. Kalicharan should not have to wait more than 5 minutes for a bus.
5. The following are the numbers of tickets issued by two policemen on 20 days:

   By first policeman: 7, 10, 14, 12, 6, 9, 11, 13, 7, 6, 10, 8, 14, 8, 12, 11, 9, 8, 10 and 15.

   By second policeman: 10, 13, 14, 11, 10, 7, 15, 11, 10, 9, 8, 12, 16, 10, 10, 14, 10, 12, 8 and 14.

   Use the sign test at 1% level of significance to test the null hypothesis that on the average the two policemen issue equal number of tickets against the alternative hypothesis that on the average the second policeman issues more tickets than the first one.
6. (a) Under what circumstances is the Fisher-Irwin test used? Explain. What is the main limitation of this test?

   (b) A housing contractor plans to build a large number of brick homes in the coming year. Two brick manufacturing concerns have given him nearly identical rates for supplying the bricks. But before placing his order, he wants to apply a test at 5% level of significance. The nature of the test is to subject each sampled brick to a force of 900 pounds. The test is performed on 8 bricks randomly chosen from a day's production of concern *A* and on the same number of bricks randomly chosen from a day's production of concern *B*. The results were as follows:

   Of the 8 bricks from concern *A*, two were broken and of the 8 bricks from concern *B*, five were broken.

   On the basis of these test results, determine whether the contractor should place order with concern *A* or with concern *B* if he prefers significantly stronger bricks.
7. Suppose that the breaking test described in problem 6(b) above is modified so that each brick is subjected to an increasing force until it breaks. The force applied at the time the brick breaks (calling it the breaking point) is recorded as under:

	Breaking-points							
Bricks of concern A	880,	950,	990,	975	895,	1030,	1025,	1010
Bricks of concern B	915,	790,	905,	900,	890,	825,	810	885.

On the basis of the above test results, determine whether the contractor should place order for bricks with concern *A* or with concern *B* (You should answer using *U* test or Wilcoxon-Mann-Whitney test).

8. The following are the kilometres per gallon which a test driver got for ten tankfuls each of three kinds of gasoline:

Gasoline *A*	30,	41,	34,	43,	33,	34,	38,	26,	29,	36
Gasoline *B*	39,	28,	39,	29,	30,	31,	44,	43,	40,	33
Gasoline *C*	29,	41,	26,	36,	41,	43,	38,	38,	35,	40.

Use the Kruskal-Wallis test at the level of significance $\alpha = 0.05$ to test the null hypothesis that there is no difference in the average kilometre yield of the three types of gasoline.

9. (a) The following are the number of students absent from a college on 24 consecutive days:
29, 25, 31, 28, 30, 28, 33, 31, 35, 29, 31, 33, 35, 28, 36, 30, 33 26, 30, 28, 32, 31, 38 and 27. Test for randomness at 1% level of significance.

(b) The following arrangement indicates whether 25 consecutive persons interviewed by a social scientist are for (*F*) or against (*A*) an increase in the number of crimes in a certain locality:
F, F, F, F, F, F, A, F, F, F, F, F, A, F, F, F, F, A, A, F, F, F, F, F, F.
Test whether this arrangement of *A*'*s* and *F*'*s* may be regarded as random at 5% as well as at 10% level of significance.

10. Use a rank correlation at the 1% significance level and determine if there is significant positive correlation between the two samples on the basis of the following information:

Blender model	*A*1	*A*2	*A*3	*B*	*C*1	*C*2	*D*1	*D*2	*E*	*F*1	*F*2	*G*1	*G*2	*H*
Sample 1	1	11	12	2	13	10	3	4	14	5	6	9	7	8
Sample 2	4	12	11	2	13	10	1	3	14	8	6	5	9	7

11. Three interviewers rank-order a group of 10 applicants as follows:

Interviewers	*Applicants*									
	a	*b*	*c*	*d*	*e*	*f*	*g*	*h*	*i*	*j*
A	1	2	3	4	5	6	7	8	9	10
B	2	3	4	5	1	7	6	9	8	10
C	5	4	1	2	3	6	7	10	9	8

Compute the coefficient of concordance (*W*) and verify the same by using the relationship between average of Spearman's *r*'s and the coefficient of concordance. Test the significance of *W* at 5% and 1% levels of significance and state what should be inferred from the same. Also point out the best estimate of true rankings.

12. Given are the values of Spearman's *r*'s as under:

$$r_{ab} = 0.607$$
$$r_{ac} = 0.429$$
$$r_{bc} = 0.393$$

Calculate Kendall's coefficient of concordance *W* from the above information and test its significance at 5% level.