Chapter 4

Non-Parametric Methods

4.1 Introduction

So far, the statistical procedures we have discussed have dealt with particular population parameters and made use of specific assumptions about the probability distributions of sample estimators, or assumptions about the nature of the population. In particular, we have usually assumed a normal population. Such tests which deals with the parameters of the population are known as parametric tests.

But, the statistical procedures described here, do not require stringent assumptions. Many of the tests do not deal with particular parameters of the population and thus are nonparametric. i-e non-parametric test doesnot make any assumption regarding the form of the parent population. Therefore, nonparametric methods are often called distribution-free methods.

In short, we define a nonparametric method is one that satisfies at least one of the following criteria.

1. The method deals with enumerative data

   (data that are frequency counts)
2. The method does not deal with specific population parameters such as μ or σ.
3. The method does not require assumptions about specific population distributions.

   (in particular, the assumption of normality).

The assumptions associated with non-parametric tests are

1. Sample observations are independent
2. The variable under study is continuous
3. Lower order moments exist.

Advantages and Disadvantages of Non-Parametric tests

Advantages

1. They do not require us to make the assumption that a population is distributed in the shape of a normal curve or another specific shape.
2. Generally they are easier to do and to understand.
3. Sometimes even formal ordering or ranking is not required.

Disadvantages

1. They ignore a certain amount of information
2. They are often not as efficient or sharp as parametric tests.
3. The non-parametric tests cannot be used to estimate parameters in the population (or) the confidence intervals for such parameters
4. It is not possible to solve certain statistical problems by using non-parametric tests. A good example is the type of problem dealt in the analysis of variance.

4.2 Uses of Non-Parametric Methods

There are four important situations in which the use of a distribution free or non-parametric technique is indicated.

1. When quick or preliminary data analysis is needed.
2. When the assumptions of a competing distribution-tied or parametric procedure are not satisfied and the consequences of this are either unknown or known to be serious.
3. When data are only roughly scaled, for example when only comparative rather than absolute magnitudes are available.
4. When the basic question of interest if distribution-free or non-parametric in nature. For example are these two samples drawn from populations with identical distributions?

We now discuss the following non-parametric tests.

1. Sign Test for Paired Data
2. Rank sum tests
   (a) Mann-whitney U-test
   (b) Kruskal-Wallis Test or H-test
3. One sample run test
4. Rank correlation test.

4.3 Sign Test

Sign test is conducted under the following circumstances.

1. When there are pair of observations on two things being compared
2. For any given pair, each of the two observations is made under similar conditions
3. No assumptions are made regarding the parent population.

Sign test is based on the signs (plus or minus) of the deviations $d_i = x_i - y_i$ not on their numerical magnitude. It is applicable only in situations where ties or zero differences between the paired observations cannot occur. When ties or zero differences are occurred, they must be excluded from the analysis and the number of paired observations is correspondingly reduced. It can be applied to a sample of paired observations and also to a sample of single observations.

4.3.1 The sign test for paired data

Let $(x_1, y_1),\ (x_2, y_2) \cdots\cdots (x_n, y_n)$ are a sample of paired observations on two random variables x and y. Let $d_i = x_i - y_i\ ;\ i = 1, 2 \cdots\cdots n.$ Measurements are such that the deviations d_i can be expressed in terms of positives or negative signs and also $d_i's$ are independent.

We take the two hypothesis as:

Null Hypothesis: $H_0 : p = 0.5$

Alternative Hypothesis: $H_1 : p \neq 0.5$

If you look carefully at the above two hypothesis, you will see that the situation is similar to the binomial distribution case. If we toss a fair coin 30 times, then probability 'p' would be 0.5 and we would expect about fifteen heads and fifteen tails. In that case, we would use the binomial distribution as the appropriate sampling distribution. We also know that when np and nq are each atleast 5, we can use the normal distribution to approximate the binomial. Thus, we can apply normal test. Here the standard error of the proportion p is given by

$$\sigma_p = \sqrt{\frac{pq}{n}}$$

The two limits of acceptance region at 5% level of significance are

$$p + 1.96\,\sigma_p \quad \text{and} \quad p - 1.96\,\sigma_p$$

It is important to note that if both np and nq are not greater than 5, then we must use the binomial distribution instead of normal distribution test.

Working Rule

1. Omitting zero differences, find the number of positive deviations in

   $d_i = x_i - y_i$. Let it be k

   When $n \leq 30$

2. Find $p' = P(u \leqslant k) = \left(\frac{1}{2}\right)^n \sum\limits_{x=0}^{k} \binom{n}{x}$ if k is the number of positive deviations

   $p' = P(u \geqslant k) = \left(\frac{1}{2}\right)^n \sum\limits_{x=k}^{n} \binom{n}{x}$ if k is the number of negative deviations.

3. If $p' \leq 0.05$, reject the null hypothesis at 5% level and accept H_0 if $p' > 0.05$.

   When $n > 30$

   Find $Z = \dfrac{u - \frac{n}{2}}{\sqrt{\frac{n}{4}}} \sim N(0, 1)$

   Where $E(u) = \text{Mean} = np = n.\dfrac{1}{2} = \dfrac{n}{2}$

   and $V(u) = npq = n.\dfrac{1}{2} \cdot \dfrac{1}{2} = \dfrac{n}{4}$.

   and $u =$ the number of negative deviations.

   $n =$ number of given items.

   If $|Z| \leq 1.96$, we accept H_0 at 5% level of significance, otherwise reject H_0.

   If $|Z| \leq 2.58$, we accept H_0 at 1% level of significance, otherwise reject H_0 if $|Z| > 2.58$.

Solved Problem 4.1

Use the sign test to see if there is a difference between the number of days required to collect on account receivable before and after a new collection policy. Use the 0.05 significance level.

Before:	33	36	41	32	39	47	34	29	32	34	40	42	33	36	27
After:	35	29	38	34	37	47	36	32	30	34	41	38	37	35	28

Solution

Null Hypothesis: $H_0 : p = 0.5$, i.e there is no difference between the two types of collections.

Alternative Hypothesis: $H_1 : p \neq 0.5$, there is a significant difference between the two types of collections.

Level of significance: $\alpha = 0.05$

Now, the sign of scores are (before - after)

$$d_i : \; - \; + \; + \; - \; + \; 0 \; - \; - \; + \; 0 \; - \; + \; - \; + \; -$$

By omitting the zero difference, we have $n = 13$
Number of '$+$' signs = 6
Number of '$-$' signs = 7

$\therefore$ Proporation : $p = \dfrac{6}{13} = 0.46$; $q = 1 - p = 0.54$

Here $np = 13 \times \dfrac{6}{13} = 6$ and $nq = 13 \times \dfrac{7}{13} = 7$ are greater than 5, we use normal distribution to approximate the binomial.

The value $Z_\propto$ for two tailed test at 5% level is 1.96

$$\therefore \text{ the standard error of } p = \sigma_p = \sqrt{\frac{pq}{n}}$$

$$= \sqrt{\frac{0.5 \times 0.5}{13}} = 0.142$$

The Limits for acceptance region

$$p + 1.96(0.142),\; p - 1.96(0.142)$$

$$\text{i.e } (0.78,\; 0.22)$$

Conclusion

The two limits of acceptance region are 0.78 and 0.22. The sample proportion $p = 0.46$ lies within these two limits, so use accept the null hypothesis and conclude that there is no difference between the two types of collections at 0.05 level of significance.

Solved Problem 4.2

The following data show the employee's rates of defective work before and after a change in the wage incentive plan. Compare the following two sets of data to see whether the charge lowered the defective units produced. Using the sign test with $\alpha = 0.01$.

Before:	8	7	6	9	7	10	8	6	5	8	10	8
After:	6	5	8	6	9	8	10	7	5	6	9	8

Solution

Null Hypothesis: $H_0 : p = 0.5$
Alternative Hypothesis: $H_1 : p < 0.5$, (one-tailed test)
Level of significance: $\alpha = 0.01$

Test Statistic

From the given data,

$$d_i : \; - \; - \; + \; - \; + \; - \; + \; + \; 0 \; - \; - \; 0$$

Here $n = 4 + 6 = 10$

$k =$ number of negative deviations = 6

Now,

$$p' = P(u \geq k) = \left(\frac{1}{2}\right)^n \sum_{x=k}^{n} \binom{n}{x} \qquad (\because\; np < 5)$$

$$= \left(\frac{1}{2}\right)^{10} \sum_{x=6}^{10} \binom{10}{x}$$

$$= \left(\frac{1}{2}\right)^{10} \left[\binom{10}{6} + \binom{10}{7} + \cdots + \binom{10}{10}\right]$$

$$= (0.000976)\,(386)$$

$$= 0.3767$$

Conclusion:

Since $p' > 0.05$, we accept the null hypothesis and conclude that there is no significant change in the defective units produced.

Solved Problem 4.3 ***(Anna University. MBA Jan.2007)***

A consumer panel includes 14 individuals. It is asked to rate two brands of cococola according to a point evaluation system based on several criteria. The table gives below reports the points assigned. Test the null hypothesis that there is no difference in the level of ratings for the two brands of cola at 5% level of significance using the sign test.

Panel member	Brand I	Brand II
1	20	16
2	24	26
3	28	18
4	24	17
5	20	20