UNIT 1 RATIONALE FOR NON-PARAMETRIC STATISTICS

Structure

1.0 INTRODUCTION

Statistics is of great importance in the field of psychology. The human behaviour which is so unpredictable and cannot be so easily measured or quantified, through statistics attempts are made to quantify the same. The manner in which one could measure human behaviour is through normal distribution concept wherein it is assumed that most behaviours are by and large common to all and only a very small percentage is in either of the extremes of normal distribution curve. Keeping this as the frame of reference, the behaviour of the individual is seen and compared with this distribution. For analysis of obtained information about human behaviour we use both parametric and non-parametric statistics. Parametric statistics require normal distribution assumptions whereas non-parametric statistics does not require these assumptions and need not also be compared with normal curve. In this unit we will be dealing with non-parametric statistics, its role and functions and its typical characteristics and the various types of non-parametric statistics that can be used in the analysis of the data.

1.1 OBJECTIVES

After reading this unit, you will be able to:

- Define non-parametric statistics;
- Differentiate between parametric and non-parametric statistics;
- Elucidate the assumptions in non-parametric statistics;
- Describe the characteristics of non-parametric statistics; and
- Analyse the use of non-parametric statistics.

1.2 DEFINITION OF NON-PARAMETRIC STATISTICS

Non-parametric statistics covers techniques that do not rely on data belonging to any particular distribution. These include (i) distribution free methods (ii) non structural models. Distribution free means its interpretation does not depend on any parametrized distributions. It deals with statistics based on the ranks of observations and not necessarily on scores obtained by interval or ratio scales.

Non-parametric statistics is defined to be a function on a sample that has no dependency on a parameter. The interpretation does not depend on the population fitting any parametrized distributions. Statistics based on the ranks of observations are one example of such statistics and these play a central role in many non-parametric approaches.

Non-parametric techniques do not assume that the structure of a model is fixed. Typically, the model grows in size to accommodate the complexity of the data. In these techniques, individual variables are typically assumed to belong to parametric distributions, and assumptions about the types of connections among variables are also made.

Non-parametric methods are widely used for studying populations that are based on rank order (such as movie reviews receiving one to four stars). The use of non-parametric methods may be necessary when data have a ranking but no clear numerical interpretation. For instance when we try to assess preferences of the individuals, (e.g. I prefer Red more than White colour etc.), we use non parametic methods. Also when our data is based on measurement by ordinal scale, we use non-parametric statistics.

As non-parametric methods make fewer assumptions, their applicability is much wider than those of parametric methods. Another justification for the use of non-parametric methods is its simplicity. In certain cases, even when the use of parametric methods is justified, non-parametric methods may be easier to use. Due to the simplicity and greater robustness, non-parametric methods are seen by some statisticians as leaving less room for improper use and misunderstanding.

A statistic refers to the characteristics of a sample, such as the average score known as the mean. A parameter, on the other hand, refers to the characteristic of a population such as the average of a whole population. A statistic can be employed for either descriptive or inferential purposes and one can use either of the two types of statistical tests, viz., parametric Tests and non-parametric Tests (assumption free test).

The distinction employed between parametric and non-parametric test is primarily based on the level of measurement represented by the data that are being analysed. As a general rule, inferential statistical tests that evaluate categorical / nominal data and ordinal rank order data are categorised as non-parametric tests, while those tests that evaluate interval data or ratio data are categorised as parametric tests.

Differences between parametric and non-parametric statistics

The parametric and non-parametric statistics differ from each other on these various levels

Level of Differences	Parametric	Non Parametric
Assumed Distribution	Normal	Any
Assumed Variance	Homogeneous	Homogenous and Heterogeneous both
Typical data	Ratio or Interval	Ordinal or Nominal
Usual Central measure	Mean	Median
Benefits	Can Draw more conclusions	Simple and less affected by extreme score

Level of measurement is an important criterion to distinguish between the parametric and non-parametric tests. Its usage provides a reasonably simple and straightforward schema for categorisation that facilitates the decision making process for selecting an appropriate statistical test.

1.3 ASSUMPTIONS OF PARAMETRIC AND NON-PARAMETRIC STATISTICS

Assumptions to be met for the use of parametric tests are given below:

- Normal distribution of the dependent variable
- A certain level of measurement: Interval data
- Adequate sample size (>30 recommended per group)
- An independence of observations, except with paired data
- Observations for the dependent variable have been randomly drawn
- Equal variance among sample populations
- Hypotheses usually made about numerical values, especially the mean

Assumptions of Non-parametric Statistics test are fewer than that of the parametric tests and these are given below.:

- An independence of observations, except with paired data
- Continuity of variable under study

Characteristics of non-parametric techniques:

- Fewer assumptions regarding the population distribution
- Sample sizes are often less stringent
- Measurement level may be nominal or ordinal
- Independence of randomly selected observations, except when paired
- Primary focus is on the rank ordering or frequencies of data
- Hypotheses are posed regarding ranks, medians, or frequencies of data

There are three major parametric assumptions, which are, and will continue to be routinely violated by research in psychology: level of measurement, sample size, and normal distribution of the dependent variable. The following sections will discuss these assumptions, and elucidate why much of the data procured in health science research violate these assumptions, thus implicating the use of non-parametric techniques.

Self Assessment Questions

1) Define Non-parametric statistics.

....

....

....

....

2) What are the charateristic featurs of non-parametric statistics?

....

....

....

....

3) Differentiate between parametric and non-parametric statistics.

....

....

....

....

4) What are the assumptions underlying parametric and non-parametric statistics?

....

....

....

....

When statistical tests are to be used one must know the following:

1.3.1 Level of Measurement

When deciding which statistical test to use, it is important to identify the level of measurement associated with the dependent variable of interest. Generally, for the use of a parametric test, a minimum of interval level measurement is required. Non-parametric techniques can be used with all levels of measurement, and are most frequently associated with nominal and ordinal level data.

1.3.2 Nominal Data

The first level of measurement is nominal, or categorical. Nominal scales are usually composed of two mutually exclusive named categories with no implied ordering: yes or no, male or female. Data are placed in one of the categories, and the numbers in each category are counted (also known as frequencies). The key to nominal level measurement is that there are no numerical values assigned to the variables. Given that no ordering or meaningful numerical distances between numbers exist in nominal measurement, we cannot obtain the coveted 'normal distribution' of the dependent variable. Descriptive research in the health sciences would make use of the nominal scale often when collecting demographic data about target populations (i.e. pain present or not present, agree or disagree).

Example of an item using a nominal level measurement scale

1) Does your back problem affect your employment status? □ Yes □ No

2) Are you limited in how many minutes you are able to walk continuously with or without support (i.e. cane)? □ Yes □ No

1.3.3 Ordinal Data

The second level of measurement, which is also frequently associated with non-parametric statistics, is the ordinal scale (also known as rank-order). Ordinal level measurement gives us a quantitative 'order' of variables, in mutually exclusive categories, but no indication as to the value of the differences between the positions (squash ladders, army ranks). As such, the difference between positions in the ordered scale cannot be assumed to be equal. Examples of ordinal scales in health science research include pain scales, stress scales, and functional scales. One could estimate that someone with a score of 5 is in more pain, more stressed, or more functional than someone with a score of 3, but not by *how much*. There are a number of non-parametric techniques available to test hypotheses about differences between groups and relationships among variables, as well as descriptive statistics relying on rank ordering. Table below provides an example of an ordinal level item from the Oswestry Disability Index.

Table: Walking (Intensity of pain in terms of the ability to walk)

S. No.	Description	Intensity
1	Pain does not prevent me walking any distance	Lowest intensity of pain
2	Pain prevents me from walking more than 2 kilometres	Some level of intensity of pain
3	Pain prevents me from walking more than 1 kilometre	Moderate intensity of pain
4	Pain prevents me from walking more than 500 meters	High intensity of pain
5	I can only walk using a stick or crutches	Very high intensity of pain

1.3.4 Interval and Ratio Data

Interval level data is usually a minimum requirement for the use of parametric techniques. This type of data is also ordered into mutually exclusive categories, but in this case the divisions between categories are equidistant. The only difference between interval data and ratio data, is the presence of a meaningful zero point. In interval level measurement, zero does not represent the absence of value. As such, you cannot say that one point is two times larger than another. For example, 100 degrees Celsius is not two times hotter than 50 degrees because zero does not represent the complete absence of heat.

Ratio is the highest level of measurement and provides the most information. The level of measurement is characterised by equal intervals between variables, and a meaningful zero point. Examples of ratio level measurement include weight, blood pressure, and force. It is important to note that in health science research we often use multi item scales, with individual items being either nominal or ordinal.

1.3.5 Sample Size

Adequate sample size is another of the assumptions underlying parametric tests. In a large number of research studies, we do use small sample size and in certain cases we just use one case study and observe that case over a period of time. Some times, we take small sample sizes from a certain place and such samples are called as convenience samples, and limited funding. Thus, the assumption of large sample size is often violated by such studies using parametric statistical techniques.

The sample size required for a study has implications for both choices of statistical techniques and resulting power. It has been shown that sample size is directly related to researchers' ability to correctly reject a null hypothesis (power). As such, small sample sizes often reduce power and increase the chance of a type II error. It has been found that by using non-parametric techniques with small sample sizes, it is possible to gain adequate power. However, there does not seem to be a consensus among statisticians regarding what constitutes a small sample size. Many statisticians argue that if the sample size is very small, there may be no alternative to using a non-parametric statistical test, but the value of 'very small' is not delineated. It has been suggested by Wampold et al. (1990) , that the issue of sample size is closely related to the distribution of the dependent variable, given that as sample size increases, the sampling distribution approaches normal(n>100).

At the same time, one can state that if the distribution of the dependant variable resembles closely the normal distribution, then it will amount to the sampling distribution of the mean being approximately normal. For other distributions, 30 observations might be required. Furthermore, in regard to decision about the statistical technique to be used, there is no clear cut choice but one can choose a technique depending on the nature of the data and sample size. Thus even the choice of parametric or non-parametric tests 'depends' on the nature of the data, the sample size, level of measurement, the researcher's knowledge of the variables' distribution in the population, and the shape of the distribution of the variable of interest. If in doubt, the researcher should try using both parametric and non-parametric techniques.

1.3.6 Normality of the Data

According to Pett (1997), in choosing a test we must consider the shape of the distribution of the variable of interest. In order to use a parametric test, we must assume a normal distribution of the dependent variable. However, in real research situations things do not come packaged with labels detailing the characteristics of the population of origin. Sometimes it is feasible to base assumptions of population distributions on empirical evidence, or past experience. However, often sample sizes are too small, or experience too limited to make any reasonable assumptions about the population parameters. Generally in practice, one is only able to say that a sample appears to come from say, a skewed, very peaked, or very flat population. Even when one has precise measurement (ratio scale), it may be irrational to assume a normal distribution, because this implies a certain degree of symmetry and spread.

Non-parametric statistics are designed to be used when we know nothing about the distribution of the variable of interest. Thus, we can apply non-parametric techniques to data from which the variable of interest does not belong to any specified distribution (i.e. normal distribution). Although there are many variables in existence that are normally distributed, such as weight, height and strength, this is not true of all variables in social or health sciences.

The incidence of rare disease and low prevalence conditions are both non-normally distributed populations. However, it seems that most researchers using parametric statistics often just 'assume' normality. Micceri et al. (1989) states that the naïve assumption of normality appears to characterise research in many fields.

However, empirical studies have documented non normal distributions in literature from a variety of fields. Micceri et al. (1989) investigated the distribution in 440 large sample achievement and psychometric measures. It was found that all of the samples were significantly non-normal (p<0.01).

It was concluded that the underlying tenets of normality assuming statistics appeared to be fallacious for the commonly used data in these samples. It is likely that if a similar study, investigating the nature of the distributions of data were to be conducted with some of the measures commonly used in health science research, a similar result would ensue, given that not all variables are normally distributed.

Self Assessment Questions

1) What are the aspects to be kept in mind before we decide to apply parametric or non-parametric tests?

....

....

....

2) What is ordinal data? Give suitable examples?

....

....

....

3) What are interval and ratio data ? Give examples.

....

....

....

4) Why is sample size important to decide about usingparametric or non-parametric tests?

....

....

....

5) What is meant by normality of a data? Explain.

....

....

....

1.4 THE USE OF NON-PARAMETRIC TESTS

It is apparent that there are a number of factors involved in choosing whether or not to use a non-parametric test, including level of measurement, sample size and sample distribution. When the choice of statistical technique for a set of data is not clear, there is no harm in analysing the data with both these methods, viz., parametric and on parametric methods.

It must be remembered that for each of the main parametric techniques there is a non-parametric test available. Also, experiments with the data would also determine which test provides the best power, and the greatest level of significance. In general, these tests fall into the following categories:

- Tests of differences between groups (independent samples);
- Tests of differences between variables (dependent samples);
- Tests of relationships between variables.

1.4.1 Differences between Independent Groups

Usually, when we have two samples that we want to compare concerning their mean value for some variable of interest, we would use the *t*-test for independent samples). The non-parametric alternatives for this test are the Wald-Wolfowitz runs test, the Mann-Whitney U test, and the Kolmogorov-Smirnov two-sample test.

If we have multiple groups, we would use analysis of variance (see ANOVA/MANOVA. The non-parametric equivalents to this method are the KruskalWallis analysis of ranks and the Median test.

1.4.2 Differences between Dependent Groups

If we want to compare two variables measured in the same sample we would customarily use the t-test for dependent samples. For example, we want to compare the math skills of students just at the beginning of the year and again at the end of the year, we would take the scores and use the t-test for such comparison and state that there is a significant difference between the two periods. Non-parametric alternatives to this test are the *Sign* test and *Wilcoxon's matched pairs* test.

If the variables of interest are dichotomous in nature (i.e., "pass" vs. "no pass") then McNemar's Chi-square test is appropriate.

If there are more than two variables that were measured in the same sample, then we would customarily use repeated measures ANOVA.

Non-parametric alternatives to this method are Friedman's two-way analysis of variance and Cochran Q test (if the variable was measured in terms of categories, e.g., "passed" vs. "failed"). Cochran Q is particularly useful for measuring changes in frequencies (proportions) across time.

1.4.3 Relationships between Variables

To express a relationship between two variables one usually computes the correlation coefficient. Non-parametric equivalents to the standard correlation coefficient of Pearson 'r' are Spearman R, Kendall Tau.

The appropriate non-parametric statistics for testing the relationship between two variables are the Chi-square test, the Phi coefficient, and the Fisher exact test. In addition, a simultaneous test for relationships between multiple cases is available, as for example, Kendall coefficient of concordance. This test is often used for expressing inter rater agreement among independent judges who are rating (ranking) the same stimuli.

1.4.4 Descriptive Statistics

When one's data are not normally distributed, and the measurements at best contain rank order information, then using non prametric methods is the best. For example, in the area of psychometrics it is well known that the rated intensity of a stimulus (e.g., perceived brightness of a light) is often a logarithmic function of the actual intensity of the stimulus (brightness as measured in objective units of Lux). In this example, the simple mean rating (sum of ratings divided by the number of stimuli) is not an adequate summary of the average actual intensity of the stimuli. (In this example, one would probably rather compute the geometric mean.) Non-parametrics and Distributions will compute a wide variety of measures of location (mean, median, mode, etc.) and dispersion (variance, average deviation, quartile range, etc.) to provide the "complete picture" of one's data.

There are a number advantages in using non-parametric techniques in health science research. The most important of these advantages are the generality and wide scope of non-parametric techniques. The lack of stringent assumptions associated with non-parametric tests implies that there is little probability of violating assumptions, which implies robustness. The application of non-parametric tests in social and Health science researcsh is wide, given that they can be applied to constructs for which it is impossible to obtain quantitative measures (descriptive studies), as well as to small sample sizes.

1.4.5 Problems and Non-parametric Tests

The most common non-parametric tests used for four different problems include the following:

i) **Two or more independent groups:** The Mann-Whitney 'U' test and the Kruskal-Wallis one-way analysis of variance (H) provide tests of the null hypothesis that independent samples from two or more groups come from identical populations. Multiple comparisons are available by the Kruskal- Wallis test.

ii) **Paired observations:** The sign test and Wilcoxon Signed-rank test both test the hypothesis of no difference between paired observations.

iii) **Randomized blocks:** The Friedman two-way analysis of variance is the non-parametric equivalent of a two-way ANOVA with one observation per cell or a repeated measures design with a single group. Multiple comparisons are available for the Friedman test. Kendall's coefficient of concordanceis a normalization of the Friedman statistic.

iv) **Rank correlations:** The Kendall and Spearman rank correlations estimate the correlation between two variables based on the ranks of the observations.

The Table below gives an overview of when to use which test:

Choosing	TEST	
	PARAMETRIC	NON PARAMETRIC
Correlation test	Pearson	Spearman
Independent Measures, 2 Groups	Independent- Measures t-test	Mann-Whitney test ('U' Test)
Independent Measures, > 2 Groups	One Way Independent Measures ANOVA	Kruskal-Wallis Test
Repeated Measures, 2 Conditions	Matched-Pair t-Test	Wilcoxon test
Repeated Measures, > 2 Conditions	One-Way, Repeated Measures ANOVA	Friedman's Test

These statistics are discussed in many texts, including Siegel (1956), Hollander and Wolfe (1973), Conover (1980), and Lehmann (1975). Each of these non-parametric statistics has a parallel parametric test.

Self Assessment Questions

1) When do we use the non-parametric statistics?

....

....

....

....

2) What is meant by descriptive statistics in the context of non-parametric statistics?

....

....

....

....

3) State when to use which test – parametric or non-parametric?

....

....

....

....

4) What are the four problems for which non-parametric statistics is used?

....

....

....

....

1.4.6 Non-parametric Statistics

The primary barrier to use of non-parametric tests is the misconception that they are less powerful than their parametric counterparts (power is the ability to correctly reject the null hypothesis). It has been suggested that parametric tests are almost always more powerful than non-parametric tests. These assertions are often made with no references to support them, suggesting that this falls into the realm of 'common knowledge'. Evidence to support this is not abundant, nor conclusive. Rather, on closer examination, it is found that parametric tests are more powerful than non-parametric tests only if all of the assumptions underlying the parametric test are met.

Pierce (1970) suggests that unless it has been determined that the data do comply with all of the restrictions imposed by the parametric test; the greater power of the parametric test is irrelevant. This is because 'the purpose of applied statistics is to delineate and justify the inferences that can be made within the limits of existing knowledge - that purpose is defeated if the knowledge assumed is beyond that actually possessed'. Thus, the power advantage of the parametric test does not hold when the assumptions of the parametric test are not met, when the data are in ranks, or when the non-parametric test is used with interval or ratio data.

When comparison studies have been made between parametric and non-parametric tests, the non-parametric tests are frequently as powerful as parametric, especially with smaller sample sizes. Blair et al. (1985) compared the power of the paired sample t-test (a common parametric test), to the Wilcoxon signed-ranks test (non-parametric), under various population shapes and sample sizes (n=10, 25, 50), using a simple pre-post test design. It was found that in some situations the t-test was more powerful than the Wilcoxon.

However, the Wilcoxon test was found to be the more powerful test in a greater number of situations (certain population shapes and sample sizes), especially when sample sizes were small. In addition, the power advantage of the Wilcoxon test often increased with larger sample sizes, suggesting that non-parametric techniques need not be limited to studies with small sample sizes. It was concluded that insofar as these two statistics are concerned, the often-repeated claim that parametric tests are more powerful than non-parametric test is not justified.

Generally, the rationale for using the t-test over the Wilcoxon test is that the parametric tests are more powerful under the assumption of normality. However, it was shown in this study that even under normal theory, there was little to gain, in terms of power by using the t-test as opposed to the Wilcoxon.

It was suggested by Blair that 'it is difficult to justify the use of a t-test in situations where the shape of the sampled population is unknown on the basis that a power advantage will be gained if the populations does happen to be normal'. Blair concluded by saying that 'although there were only two tests compared here, it should be viewed as part of a small but growing body of evidence that is seriously challenging the traditional views of non-parametric statistics'. This study demonstrated that the use of non-parametric techniques is implicated whenever there is doubt regarding the fulfilment of parametric assumptions, such as normality or sample size.

Self Assessment Questions

Answer the following as True or False.

1) Parametric tests are equally assumptive as Non-parametric tests. T / F

2) Non-parametric tests are most applicable when data is in rank form. T / F

3) Small sample size is not entertained by parametric tests. T / F

4) Parametric tests are more statistically grounded than Non-parametric tests. T / F

5) Non-parametric statistics cannot be used for complex research designs. T / F

1.4.7 Advantages and Disadvantages of Non-parametric Statistics

Advantages

1) Non-parametric test make less stringent demands of the data. For standard parametric procedures to be valid, certain underlying conditions or assumptions must be met, particularly for smaller sample sizes. The one-sample t test, for example, requires that the observations be drawn from a normally distributed population. For two independent samples, the t test has the additional requirement that the population standard deviations be equal. If these assumptions/conditions are violated, the resulting P-values and confidence intervals may not be trustworthy. However, normality is not required for the Wilcoxon signed rank or rank sum tests to produce valid inferences about whether the median of a symmetric population is 0 or whether two samples are drawn from the same population.

2) Non-parametric procedures can sometimes be used to get a quick answer with little calculation.

   Two of the simplest non-parametric procedures are the sign test and median test. The *sign test* can be used with paired data to test the hypothesis that differences are equally likely to be positive or negative, (or, equivalently, that the median difference is 0). For small samples, an exact test of whether the proportion of positives is 0.5 can be obtained by using a binomial distribution. For large samples, the test statistic is:

   $(\text{plus - minus})^2 / (\text{plus + minus})$, where *plus* is the number of positive values and *minus* is the number of negative values. Under the null hypothesis that the positive and negative values are equally likely, the test statistic follows the chi-square distribution with 1 degree of freedom. Whether the sample size is small or large, the sign test provides a quick test of whether two paired treatments are equally effective simply by counting the number of times each treatment is better than the other.

   Example: 15 patients given both treatments A and B to test the hypothesis that they perform equally well. If 13 patients prefer A to B and 2 patients prefer B to A, the test statistic is $(13 - 2)^2 / (13 + 2)$ [= 8.07] with a corresponding P-value of 0.0045. The null hypothesis is therefore rejected.

   The *median test* is used to test whether two samples are drawn from populations with the same median. The median of the combined data set is calculated and each original observation is classified according to its original sample (A or B) and whether it is less than or greater than the overall median. The chi-square test for homogeneity of proportions in the resulting 2-by-2 table tests whether the population medians are equal.

3) Non-parametric methods provide an air of objectivity when there is no reliable (universally recognized) underlying scale for the original data and there is some

concern that the results of standard parametric techniques would be criticized for their dependence on an artificial metric. For example, patients might be asked whether they feel *extremely uncomfortable* / *uncomfortable* / *neutral* / *comfortable* / *very comfortable*. What scores should be assigned to the comfort categories and how do we know whether the outcome would change dramatically with a slight change in scoring? Some of these concerns are blunted when the data are converted to ranks[4].

4) A historical appeal of rank tests is that it was easy to construct tables of exact critical values, provided there were no ties in the data. The same critical value could be used for all data sets with the same number of observations because every data set is reduced to the ranks 1,...,*n*. However, this advantage has been eliminated by the ready availability of personal computers.

5) Sometimes the data do not constitute a random sample from a larger population. The data in hand are all there are. Standard parametric techniques based on sampling from larger populations are no longer appropriate. Because there are no larger populations, there are no population parameters to estimate. Nevertheless, certain kinds of non-parametric procedures can be applied to such data by using *randomization models*.

From Dallal (1988): Consider, for example, a situation in which a company's workers are assigned in haphazard fashion to work in one of two buildings. After a year physical tests are administered, it appears that workers in one building have higher lead levels in their blood. Standard sampling theory techniques are inappropriate because the workers do not represent samples from a large population—there is no large population. The randomization model, however, provides a means for carrying out statistical tests in such circumstances. The model states that if there were no influence exerted by the buildings, the lead levels of the workers in each building should be no different from what one would observe after combining all of the lead values into a single data set and dividing it in two, at random, according to the number of workers in each building. The stochastic component of the model, then, exists only in the analyst's head; it is not the result of some physical process, except insofar as the haphazard assignment of workers to buildings is truly random.

Of course, randomization tests cannot be applied blindly any more than normality can automatically be assumed when performing a t test. (Perhaps, in the lead levels example, one building's workers tend to live in urban settings while the other building's workers live in rural settings. Then the randomization model would be inappropriate.) Nevertheless, there will be many situations where the less stringent requirements of the randomization test will make it the test of choice. In the context of randomization models, randomization tests are the ONLY legitimate tests; standard parametric test are valid only as approximations to randomization tests.

Disadvantages

Such a strong case has been made for the benefits of non-parametric procedures that some might ask why parametric procedures are not abandoned entirely in favour of non-parametric methods!

The major disadvantage of non-parametric techniques is contained in its name. Because the procedures are *non-parametric*, there are no parameters to describe and it becomes more difficult to make quantitative statements about the actual difference between populations. (For example, when the sign test says two treatments are different, there is no confidence interval and the test does not say by how much the treatments differ.)

However, it is sometimes possible with the right software to compute estimates (and even confidence intervals!) for medians, differences between medians. However, the calculations are often too tedious for pencil-and-paper. A computer is required. As statistical software goes though its various iterations, such confidence intervals may become readily available, but are not there.

The second disadvantage is that non-parametric procedures throw away information. The sign test, for example, uses only the signs of the observations. Ranks preserve information about the order of the data but discard the actual values. Because information is discarded, non-parametric procedures can never be as powerful (able to detect existing differences) as their parametric counterparts when parametric tests can be used.

1.5 MISCONCEPTIONS ABOUT NON-PARAMETRIC TESTS

The lack of use of non-parametric techniques is owing to a series of common misconceptions about this branch of statistics.

Non-parametric statistics have long taken the back seat to parametric statistics, often being portrayed as inferior in practice and teaching. It has been suggested that researchers are hesitant to use these techniques, due to fears that peer reviewers may not be completely familiar with these statistics, and therefore unable to properly interpret, and review the results.

The above opinion could be due to the widespread case of limited exposure of researchers and clinicians to this type of statistics.

Non-parametric techniques are often left out of basic statistics courses, and relegated to the last chapter of texts, making them seem less important, while reinforcing the focus on parametric statistics.

Another common misconception concerning non-parametric statistics is that they are restricted in their application. It is thought that there are only a limited number of simple designs that can be analysed using these techniques.

However, there are non-parametric techniques which span from simple 2-group analysis, to complex structural equation modelling. Basically, for any parametric test, there is a non-parametric equivalent that would be equally, or in some cases, more appropriate for use.

Self Assessment Questions

1) What are the advantages of non-parametric statistics?

....

....

....

2) What are the disadvantages of non-parametric statistics?

....

....

....

3) What are the misconceptions about non-parametric statistic tests?

....

....

....

1.6 LET US SUM UP

The key points of our discussion in this unit are:

1) **Characteristics common to most non-parametric techniques:**

- Fewer assumptions regarding the population distribution
- Sample sizes are often less stringent
- Measurement level may be nominal or ordinal
- Independence of randomly selected observations, except when paired
- Primary focus is on the rank ordering or frequencies of data
- Hypotheses are posed regarding ranks, medians, or frequencies of data

2) **Conditions when it is appropriate to use a non-parametric Test:**

- Nominal or ordinal level of measurement
- Small sample sizes
- Non-normal distribution of dependent variable
- Unequal variances across groups
- Data with notable outliers

3) **Advantages and disadvantages of Non-parametric Tests:**

- Methods quick and easy to apply
- Theory fairly simple
- Assumptions for tests easily satisfied
- Accommodate unusual or irregular sample distributions
- Basic data need not be actual measurements
- Use with small sample sizes
- Inherently robust due to lack of stringent assumptions
- Process of collecting data may conserve time and funds
- Often offer a selection of interchangeable methods
- Can be used with samples made up of observations from several different populations

1.7 UNIT END QUESTIONS

1) What are the major differences between parametric and non-parametric statistics?

2) Enumerate the advantages of non-parametric statistics.

3) Are there any assumptions for "Assumption Free tests"? If yes what are the assumptions of non-parametric statistics?

4) "Non-parametric Statistics has much wider scope than parametric statistics" support the statement with your arguments.

5) What are the major misconceptions regarding non-parametric statistics?

1.8 SUGGESTED READING

Cohen J. (1988) *Statistical Power Analysis for the Behavioural Sciences*. Hillsdale, NJ: Lawrence Erlbaum.

Micceri T. (1989) The Unicorn, The Normal Curve, and Other Improbable Creatures. *Psychological Bulletin*, 156-66.

Kerlinger F.N. (1964) *Foundations of Behavioural Research*. New York: Holt, Rinehart and Winston.

Pett M.A. (1997) *Non-parametric Statistics for Health Care Research*. London, Thousand Oaks, New Delhi: Sage Publications.

Siegel S. and Castellan N.J. (1988) *Non-parametric Statistics for the Behavioral Sciences* (2nd edition). New York: McGraw Hill.

Wampold BE & Drew CJ. (1990) *Theory and Application of Statistics*. New York: McGraw-Hill.

UNIT 2 MANN WHITNEY ‘U’ TEST FOR TWO SAMPLE TEST

Structure

2.0 INTRODUCTION

Non-parametric statistics are distribution free statistics and can be used for small samples as well as any kind of distribution. It has many tests which are equivalent to the parametric tests. For instance for the tests like mean, we have Mann Whitney U test, for Pearson ‘r’ we have Kendall tau test and so on. The non-parametric tests are available for single sample, matched pair sample, two samples and k samples. In this unit we will be dealing with Two sample tests and the various non-parametric tests that we can use analyse data if we have two samples. We will initially start with the definition of what is two sample test and go on to present different non-parametric statitistics that could be applied to analyse such data and then finally present how to solve problems based on such data.

2.1 OBJECTIVES

After reading this unit, you will be able to:

- Define two sample data;
- Explain what are two sample tests;

- Present the various non-parametric tests that can be used to analyse two sample data;
- Explain the significance levels and interpretation of such data; and
- Solve problems in two sample data.

2.2 DEFINITION OF TWO SAMPLE TESTS

Two sample tests are those which we call as tests of independence rather than goodness of fit tests. We are testing to see whether or not 2 variables are "related" or "dependent". Thus, the Ho , that is the Null Hypothesis, takes the general form

Ho: x and y are independent.

In a parametric test, we have seen earlier that to find out if two groups differ in performance , we used the t-test and if the t value was significant at .05 level, we rejected the null hypothesis and concluded that the two groups differed in their performance. In this type of t test we required the performance to be nornmally distributed and the sample size to be more than 30 and such other parametric test conditions. However if the sample size is less than 30 and the data is not normally distributed we would use the non-parametric test to find out if the two groups differed in their performance.

Let us say the 2 samples are males and females . We are comparing their marks in History at the final examination. Here one sample is gender, categorised into male and female. This is a nominal scale measurement. The other is 'scores obtained in History', a continuous variable which can range from zero to 100 or more depending on out of how many marks a score has been taken as performance.

Thus one variable , let us say Gender , that is X variable is in nominal scale of measurement and dichotomous (takes on only 1 of 2 possible values – that is male or female). The marks in History Y variable is treated as continuous (can take on a whole range of values on a continuum). To find out if males scored significantly higher than females in history in the final examination, we may if the sample size was more than 30, apply t-test and if the t value is significant, and also males average or the mean score is higher than that of female students, then we will conclude that males have scored significantly higher in history as compared to female students.

To take another example, let us say 2 groups matched in many respects each receive a different teaching method, and their final exam scores are compared.

Let us say X = Teaching method categorised into 1 and 2 methods.(X1 and X2)

Y = the final exam scores

Ho: X1 and X2 will not differ in terms of marks obtained

Here X = the independent variable and

Y = the dependent variable. (Marks in history)

We are trying to find out which of the two teaching methods is producing higher marks in final exam performance. The null hypothesis states there will be no difference in the marks obtained irrespective of X1 or X2 method of teaching.

Now all the subjects marks are taken for the two groups of persons undergoing two different methods of teaching. For this the ideal non-parametric test will be the t-test.

If the t-value goes beond the value given in the table at .05 level, we reject the null hypothesis and state that the two teaching methods do bring about a change in the performance of the students.

However if the sample size is small and there is no assumption of normality, then we would apply non-parametric test. These tests help in rejecting or accepting null hypothesis depending on the analysis.

Now let us see what are the various tests do we have under the non-parametrict test that can be applied.

2.3 MANN WHITNEY 'U' TEST

The Mann-Whitney (Wilcoxon) rank-sum test is a non-parametric analog of the two-sample *t* test for independent samples. The *Mann-Whitney U test* is a non-parametric test that can be used to analyse data from a two-group independent groups design when measurement is at least ordinal. It analyses the *degree of separation* (or the amount of overlap) between the Experimental (E) and Control (C) groups.

The *null hypothesis* assumes that the two sets of scores (E and C) are samples from the same population; and therefore, because sampling was random, the two sets of scores *do not differ systematically* from each other.

The *alternative hypothesis*, on the other hand, states that the two sets of scores *do* differ systematically. If the alternative is directional, or one-tailed, it further specifies the direction of the difference (i.e., Group E scores are systematically higher or lower than Group C scores).

The statistic that is calculated is either U or U'.

U1 = the number of Es less than Cs

U2 = the number of Cs less than Es

U = the smaller of the two values calculated above

U' = the larger of the two values calculated above

When you perform these tests, your data should consist of a random sample of observations from two different populations. Your goal is to compare either the location parameters (medians) or the scale parameters of the two populations. For example, suppose your data consist of the number of days in the hospital for two groups of patients: those who received a standard surgical procedure and those who received a new, experimental surgical procedure. These patients are a random sample from the population of patients who have received the two types of surgery. Your goal is to decide whether the median hospital stays differ for the two populations.

2.4 RELEVANT BACKGROUND INFORMATION ON 'U' TEST

The Mann-Whitney *U* test is employed with ordinal (rank-order) data in a hypothesis testing situation involving a design with two independent samples. If the result of the Mann-Whitney *U* test is significant, it indicates there is a significant difference between the two sample medians, and as a result of the latter the researcher can conclude there is a high likelihood that the samples represent populations with different median values.

Two versions of the test to be described under the label of the Mann-Whitney *U* test were independently developed by Mann and Whitney (1947) and Wilcoxon (1949).

The version to be described here is commonly identified as the Mann-Whitney *U* test, while the version developed by Wilcoxon (1949) is usually referred to as the Wilcoxon-Mann-Whitney test.' Although they employ different equations and different tables, the two versions of the test yield comparable results.

In employing the Mann-Whitney *U* test, one of the following is true with regard to the rank order data that are evaluated:

a) The data are in a rank order format, since it is the only format in which scores are available; or

b) The data have been transformed into a rank order format from an interval ratio format, since the researcher has reason to believe that the normality assumption (as well as, perhaps, the homogeneity of variance assumption) of the t test for two independent samples (which is the parametric analog of the Mann-Whitney U test) is saliently violated.

It should be noted that when a researcher elects to transform a set of interval/ratio data into ranks, information is sacrificed. This latter fact accounts for the reluctance among some researchers to employ non-parametric tests such as the Mann-Whitney *U* test, even if there is reason to believe that one or more of the assumptions of the t test for two independent samples have been violated.

Various sources (e.g., Conover (1980, 1999), Daniel (1990), and Marascuilo and McSweeney (1977)) note that the Mann-Whitney U test is based on the following assumptions:

a) Each sample has been randomly selected from the population it represents;

b) The two samples are independent of one another;

c) The original variable observed (which is subsequently ranked) is a continuous random variable. In truth, this assumption, which is common to many non-parametric tests, is often not adhered to, in that such tests are often employed with a dependent variable which represents a discrete random variable; and

d) The underlying distributions from which the samples are derived are identical in shape. The shapes of the underlying population distributions, however, do not have to be normal.

Maxwell and Delaney (1990) pointed out the assumption of identically shaped distributions implies equal dispersion of data within each distribution. Because of this, they note that like the *t* test for two independent samples, the Mann-Whitney *U* test also assumes homogeneity of variance with respect to the underlying population distributions.

Because the latter assumption is not generally acknowledged for the Mann-Whitney *U* test, it is not uncommon for sources to state that violation of the homogeneity of variance assumption justifies use of the Mann-Whitney *U* test in lieu of the *t* test for two independent samples.

It should be pointed out, however, that there is some empirical evidence which suggests that the sampling distribution for the Mann-Whitney *U* test is not as affected by violation of the homogeneity of variance assumption as is the sampling distribution for *t* test for two independent samples. One reason cited by various sources for employing the Mann-

Whitney *U* test is that by virtue of ranking interval/ratio data, a researcher will be able to reduce or eliminate the impact of outliers.

Self Assessment Questions

1) Which non-parametric test should we use when the data is obtained from two different samples (Independent of each other) and we wish to see the difference between the two samples on a particular variable?

....

....

....

....

2) What is the underlying assumption of Mann-Whitney U test?

....

....

....

....

2.5 STEP BY STEP PROCEDURE FOR 'U' TEST FOR SMALL SAMPLE

Step-by-step procedure

Mann Whitney U Test for Small Sample case (not more than 20 items in each set), use U if the data is

a) in the form of ranks or

b) not normally distributed

c) there is an obvious difference in the variance of the two groups.

STEP 1: Rank the data (taking both groups together) giving rank 1 to the lowest score, and the highest rank to then highest score.

STEP 2: Find the sum of the ranks for the smaller sample

STEP 3: Find the sum of the ranks for the larger sample

STEP 4: Find U applying the formula given below:

$U = N_1N_2 + [N_1(N_1 + 1) / 2] - \Sigma R_1$

and

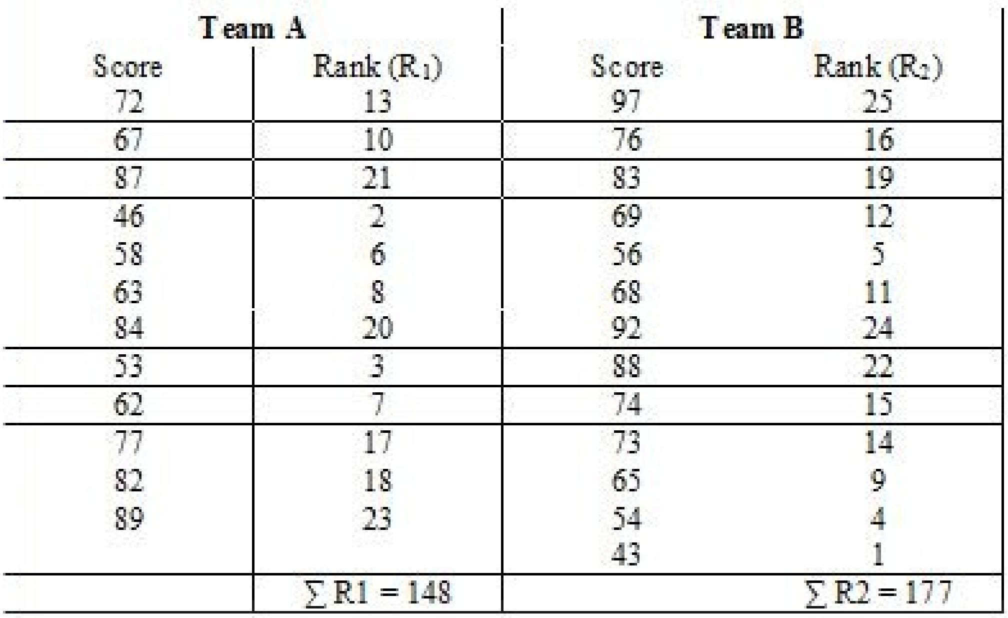
$U' = N_1N_2 + [N_2(N_2 + 1) / 2] - \Sigma R_2$

STEP 5: Look up the smaller of U and U' in Table H. There is a significant difference if the observed value is equal to or more than the table value.

STEP 6: Translate the results of the test back in the terms of experiment.

Worked Up Example:

Assessment Center Rating By Two Teams: Officers Randomly Assigned to Teams



Team A		Team B	
Score	Rank (R_1)	Score	Rank (R_2)
72	13	97	25
67	10	76	16
87	21	83	19
46	2	69	12
58	6	56	5
63	8	68	11
84	20	92	24
53	3	88	22
62	7	74	15
77	17	73	14
82	18	65	9
89	23	54	4
		43	1
	∑ R1 = 148		∑ R2 = 177

Step 1: Rank the ratings from lowest to highest regardless of assessment team.

Step 2: Sum the ranks in either group

$\Sigma (R_1) = 148$

$\Sigma (R_2) = 177$

Step 3: Calculate U

$U = N_1N_2 + [N_1(N_1 + 1) / 2] - \Sigma R_1$

$U = (12)(13) + [12 (12 + 1) / 2] - 148$

$U = 156 + 78 - 148 = 86$

And Calculate U'

$U' = N_1N_2 + [N_2(N_2 + 1) / 2] - \Sigma R_2$

$U' = (12)(13) + [13(13 + 1) / 2] - 177$

$U' = 156 + 91 - 175 = 70$

Step 4: Determine the significance of U

Decide whether you are making a one- or a two-tailed decision

Compare the smaller value of U to the appropriate critical table value for N_1 and N_2

If the observed U is smaller than the table value, the result is significant.

Step 5: The critical value of U for $N_1 = 12$ and $N_2 = 13$, two-tailed $\alpha = 0.05$, is 41.

Since the smaller obtained value of U ($U' = 70$) is larger than the table value, the null hypothesis is accepted. And we conclude that there is no significant difference in the ratings given by the two assessment teams.

2.6 STEP BY STEP PROCEDURE FOR 'U' TEST FOR LARGE SAMPLE

When both sample sizes are greater than about 20, the sampling distribution of U is for practical purposes, normal. Therefore, under these conditions, one can perform a z-test as follows:

The procedure to obtain U is similar as in small sample case (Step 1 to 3). Then the formula for Z is applied as:

$$Z = [U - (N_1N_2) / 2] / \quad N_1N_2 \sqrt{(N_1 + N_2 + 1)/12}$$

If we are dealing with a two-tailed test, then the observed z is significant at the 5 per cent level if it exceeds 1.96. For one tailed test, 5 per cent significance is attained if z exceeds 1.64 (Check these in table D in Statistics book original).

The ranking procedure can become quite laborious in large samples. Partly for this reason and partly because violation of the assumptions behind parametric statistics become less important for large sample, the Mann Whitney U test tends to be restricted to use with relatively small samples.

Self Assessment Questions

1) What unit of sample is considered as an appropriate sample for Mann Whitney U test for small sample?

....

....

....

2) What is the rationale of applying Z test be applied in a non-parametric setting?

....

....

....

2.7 COMPUTING MANN-WHITNEY U TEST IN SPSS

Step 1.Choose Analyse

Step 2. Select Non-parametric Tests

Step 3. Select 2 Independent Samples

Step 4. Highlight your test variable (in our example this would be age) and click on the arrow to move this into the Test Variable List box

Step 5. Highlight the grouping variable and click on the arrow to move this into the Grouping Variable box.

Step 6. Click on Define Groups and type in the codes that indicate which group an observation belongs to (in our example, the codes which indicate whether a subject is male or female). Click on Continue

Step 7. Under Test Type make sure that Mann-Whitney U is selected

Step 8. If you want exact probabilities, click on Exact, choose Exact, then Continue

Click on OK

2.8 WILCOXON MATCHED PAIR SIGNED RANK TEST

The Wilcoxon Matched Pair signed-ranks test is a non-parametric test that can be used for 2 repeated (or correlated) measures when measurement is at least ordinal. But unlike the sign test, it *does* take into account the magnitude of the difference.

In using this test, the difference is obtained between each of N pairs of scores observed on matched objects, for example, the difference between pretest and post-test scores for a group of students.

The difference scores obtained are then ranked.

The ranks of negative score differences are summed and the ranks of positive score differences are summed.

The test statistic T is the smaller of these two sums.

Difference scores of 0 are eliminated since a rank cannot be assigned.

If the null hypothesis of no difference between the groups of scores is true, the sum of positive ranks should not differ from the sum of negative ranks beyond that expected by chance.

2.9 RELEVANT BACKGROUND INFORMATION ON WILCOXON TEST

The Wilcoxon matched-pairs signed-ranks test (Wilcoxon (1945, 1949)) is a non-parametric procedure employed in a hypothesis testing situation involving a design with two dependent samples. Whenever one or more of the assumptions of the t test for two dependent samples are saliently violated, the Wilcoxon matched-pairs signed-ranks test (which has less stringent assumptions) may be preferred as an alternative procedure.

The Wilcoxon matched-pairs signed-ranks test is essentially an extension of the Wilcoxon signed-ranks test (which is employed for a single sample design) to a design involving two dependent samples.

In order to employ the Wilcoxon matched-pairs signed ranks test, it is required that each of n subjects (or n pairs of matched subjects) has two interval/ratio scores (each score having been obtained under one of the two experimental conditions).

A difference score is computed for each subject (or pair of matched subjects) by subtracting a subject’s score in Condition 2 from his score in Condition 1.

The hypothesis evaluated with the Wilcoxon matched-pairs signed-ranks test is whether or not in the underlying populations represented by the sampled experimental conditions, the median of the difference scores equals zero.

If a significant difference is obtained, it indicates that there is a high likelihood that the two sampled conditions represent two different populations.

The Wilcoxon matched-pairs signed-ranks test is based on the following assumptions:

a) The sample of n subjects has been randomly selected from the population it represents;

b) The original scores obtained for each of the subjects are in the format of interval/ ratio data; and

c) The distribution of the difference scores in the populations represented by the two samples is symmetric about the median of the population of difference scores.

As is the case for the t test for two dependent samples, in order for the Wilcoxon matched pairs signed ranks test to generate valid results, the following guidelines should be adhered to:

a) To control for order effects, the presentation of the two experimental conditions should be random or, if appropriate, be counterbalanced; and

b) If matched samples are employed, within each pair of matched subjects each of the subjects should be randomly assigned to one of the two experimental conditions

As is the case with the t test for two dependent samples, the Wilcoxon matched-pairs signed-ranks test can also be employed to evaluate a "one-group pretest-posttest" design. The limitations of the one group pretest posttest design are also applicable when it is evaluated with the Wilcoxon matched pairs signed ranks test.

It should be noted that all of the other tests in this text that rank data (with the exception of the Wilcoxon signed-ranks test), ranks the original interval/ratio scores of subjects.

The Wilcoxon matched-pairs signed-ranks test, however, does not rank the original interval/ratio scores, but instead ranks the interval/ratio difference scores of subjects (or matched pairs of subjects).

For this reason, some sources categorise the Wilcoxon matched-pairs signed-ranks test as a test of interval/ratio data.

Most sources, however, categorise the Wilcoxon matched-pairs signed-ranks test as a test of ordinal data, by virtue of the fact that a ranking procedure is part of the test protocol.

Self Assessment Questions

1) Which non-parametric test should we use when the data is obtained from two related sample and we wish to see the difference between the two samples on a particular variable?

....

....

....

....

2) Which one assumption does not apply to Wilcoxon Matched Pair Test, which applies to Mann Whitney U test?

....

....

....

....

3) What is the difference between t Test for Matched Pair sample and Wilcoxon Matched Pair Test?

....

....

....

....

2.10 STEP BY STEP PROCEDURE FOR WILCOXON TEST FOR SMALL SAMPLE

Step-by-step procedure

Wilcoxon Test-Small Sample case (not more than 25 pairs of scores).

For matched pairs or repeated measures designs: use instead of a correlated *t*-test if either

a) the differences between treatments can only be ranked in size or

b) the data is obviously non-normal or

c) there is an obvious difference in the variance of the two groups.

STEP 1: Obtain the difference between each pair of reading, taking sign into account

STEP 2: Rank order these differences (ignoring the sign), giving rank 1 to the smallest difference

STEP 3: Obtain *T,* the sum of the ranks for differences with the less frequent sign

STEP 4: Consult Table J. If the observed *T* is equal to or less than the table value then there is a significant difference between two conditions

STEP 5: Translate the result of the test back in terms of the experiment

Worked Up Example:

Eight pairs of twins were tested in complex reaction time situations; one member of each pair was tested after drinking 3 double whiskies, the other member was completely sober. The following reaction times were recorded:

Sober Group	Whisky Group	Step 1: Differences	Step 2:Ranks
310	300	-10	1
340	320	-20	2
290	360	70	5
270	320	50	4
370	540	170	6
330	360	30	3
320	680	360	7
320	1120	800	8

STEP 3: Less frequent sign of difference is negative,

$T = 1 + 2 = 3$

STEP 4: From Table J, when N = 8, T = 4. As the observed value of *T* is less than the table value, there is a significant difference between the two conditions.

STEP 5: Complex reaction time scores are significantly higher after drinking 3 double whiskies than when sober.

2.11 STEP BY STEP PROCEDURE FOR WILCOXON TEST FOR LARGE SAMPLE

When both sample sizes are greater than about 20, the sampling distribution of U is (for practical purposes) normal.

As with the Mann Whitney U test, the sampling distribution of the statistics (In this case *T*) approaches the normal distribution as the sample size becomes large. Therefore, under these conditions, again one can perform a z-test as follows:

$$Z= \{T- (N(N+1) / 4)\}/ N(N+1)(2N+1) / \sqrt{24}$$

The significance decisions are identical to those for the Mann Whitney largesample case. Thus, if we have a two tailed test, the observed z is significant at the 5 per cent level if it exceeds 1.96. For the one-tailed test, significance is attained if z exceeds 1.64. However, as with the Mann-Whitney test, and for the same reasons, the Wilcoxon test tends to be restricted to use with relatively small samples.

Self Assessment Questions

What unit of sample is considered as an appropriate sample for Mann Whitney U test for small sample?Give the underlying assumptions of Mann Whitney U test?

....

....

....

....

2.12 COMPUTING THE WILCOXON SIGNED RANK SUM TEST IN SPSS

- Choose Analyse
- Select Non-parametric Tests
- Select 2 Related Samples
- Specify which two variables comprise your pairs of observation by clicking on them both then clicking on the arrow to put them under Test Pair(s) List.
- Under Test Type select Wilcoxon

If you want exact probabilities (i.e. based on the binomial distribution), click on Exact, choose Exact, then Continue

Click on OK

2.13 COMPARISON OF MANN-WHITNEY 'U' TEST AND WILCOXON MPSR TEST WITH T-TEST

The power efficiency of the Mann-Whitney and Wilcoxon tests, whilst usually somewhat lower than the corresponding *t*-test, compares very favourably with it. The Mann-Whitney and Wilcoxon tests can be used in situations where the *t* –Test would be in-appropriate (e.g. where the assumptions of the *t*-test obviously do not apply). In other words, they are capable of wider application.

Different statisticians give different advice as to the relative merits of parametric and non-parametric tests. The non-parametric camp claims that their tests are simpler to compute, have fewer assumptions and can be used more widely. The parametric camp claims that their tests are robust with respect to violations of their assumptions and have greater power efficiency.

The strategy recommended here is to use the *t*-test unless the data is in form of ranks, or where the sample is small and either the distribution is obviously non-normal or there are obviously large differences in variance.

However, if you are particularly pressed for time or have a large number of analyses to do there is particularly nothing inappropriate about using non-parametric statistics, even in cases where *t*-tests might have been used.

2.14 LET US SUM UP

Two Sample test can be of two types independent sample test (two different samples being tested on one variable wherein one sample does not affect the other sample) or paired or dependent sample (same sample being tested twice or sample have some relation with each other).

The *t*-test is the parametric test for a two sample test, in non-parametric tests, Mann-Whitney 'U' test and Wilcoxon test are used for independent and paired sample respectively.

Both these tests have their own advantages, and can be used for a smaller sample size, do not have too many assumptions and can be used more widely.

2.15 UNIT END QUESTIONS

1) A researcher had an experimental group of m = 3 cases and a control group of n = 4 cases. The scores were as following:

   Exprimental scores: 9, 11, 15

   Control scores: 6, 8, 10, 13

2) In the problem 1 above, Assume these groups are independent, apply appropriate statistics and state whether the experimental condition and control conditions differ or not.

3) In the problem 1 above, Assume these groups are correlated, apply appropriate statistics and state whether the experimental condition and control conditions differ or not.

4) Doctor Radical, a math instructor at Logarithm University, has two classes in advanced calculus. There are six students in Class 1 and seven students in Class 2.

The instructor uses a programmed textbook in Class 1 and a conventional textbook in Class 2. At the end of the semester, in order to determine if the type of text employed influences student performance, Dr. Radical has another math instructor, Dr. Root, to rank the 13 students in the two classes with respect to math ability. The rankings of the students in the two classes follow:

Class 1: 1, 3, 5, 7, 11, 13

Class 2: 2, 4, 6, 8, 9, 10, 12

(Assume the lower the rank the better the student).

5) To 4 above, Apply appropriate statistics and tell if the type of text employed influenced students performance?

6) Why should you not use the large-sample *z*-test version of a non-parametric test when you have samples small enough to allow the use of small sample version?

7) Identify the non-parametric test that ought to be used.

8) You have 5 independent groups of subjects, with different numbers per group. There is also substantial departure from homogeneity of variance. The null hypothesis states that there are no differences between the groups.

You have the same situation described in question 4 (a); and in addition, the alternative hypothesis states that when the mean ranks for the 5 groups are listed from smallest to largest, they will appear in a particular *pre-specified* order.

2.16 SUGGESTED READINGS

Daniel, W. W. (1990) *Applied Non-parametric Statistics*, 2d ed. Boston: PWS-Kent.

Johnson, Morrell, and Schick (1992), Two-Sample Non-parametric Estimation and Confidence Intervals Under Truncation, *Biometrics*, 48, 1043-1056.

Siegel S. and Castellan N.J. (1988) *Non-parametric Statistics for the Behavioral Sciences* (2nd edition). New York: McGraw Hill.

Wampold BE & Drew CJ. (1990) *Theory and Application of Statistics*. New York: McGraw-Hill.

UNIT 3 KRUSKAL WALLIS ANALYSIS OF VARIANCE

Structure

3.0 INTRODUCTION

So far in Unit 2 we have studied appropriate statistical tests when we wish to compare two groups (t test if data is from a normal population, Mann-Whitney U test or Wilcoxon test if there are no assumptions about the distribution of the data), but what if there are more than two groups that require comparison? One may think that we may apply the same tests in that condition too. Like for example, if there are three groups say A, B, and C, one may see the difference between A&B, B&C and A&C. This may not look so cumbersome. Now, think if we need to compare 5 groups, A, B, C, D, E, the number for comparison tests we need to do would be 10 (A&B, A&C, A&D, A&E, B&C, B&D, B&E, C&D, C&E, D&E). And what if we need to compare 6 groups? Number of two sample test in these cases become too cumbersome and may not be feasible at all. This may further lead to unnecessary calculations and also give rise to type I error. The answer in these cases when we have more than two groups (>2 groups) to be compared is to conduct Analysis of Variance.

3.0 OBJECTIVES

After reading this unit, you will be able to:

- Define ANOVA tests;
- Describe the procedure for ANOVA calculations;
- Explain Kruskal Wallis ANOVA;
- Enumerate the conditions when this test can be applied; and
- Analyse Kruskal Wallis Anova with one way ANOVA of parametric test.

3.2 ANALYSIS OF VARIANCE

The term analysis of variance (for which the acronym ANOVA is often employed) describes a group of inferential statistical procedures developed by the British statistician Sir Ronald Fisher. Analysis of variance is all about examining the amount of variability in a *y* (response) variable and trying to understand where that variability is coming from. One way that you can use ANOVA is to compare several populations regarding some quantitative variable, *y*. The populations you want to compare constitute different groups (denoted by an *x* variable), such as political affiliations, age groups, or different brands of a product. ANOVA is also particularly suitable for situations involving an experiment where you apply certain treatments *(x)* to subjects, and you measure a response *(y)*.

Null hypothesis H_O%: Population means are equal. There will be no difference in the population means.

ì1 = ì2 = ì3 = ì4

Alternative hypothesis: *h*1

Population means are not equal. There will be difference in the means of the different populations.

The logic used in ANOVA to compare means of multiple groups is similar to that used with the t-test to compare means of two independent groups. When one way ANOVA is applied to the special case of two groups, this one way ANOVA gives identical results as the t-test.

Not surprisingly, the assumptions needed for the t-test are also needed for ANOVA. We need to assume:

1) random, independent sampling from the k populations;

2) normal population distributions;

3) equal variances within the k populations.

Assumption 1 is crucial for any inferential statistic. As with the t-test, Assumptions 2 and 3 can be relaxed when large samples are used, and Assumption 3 can be relaxed when the sample sizes are roughly the same for each group even for small samples. (If there are extreme outliers or errors in the data, we need to deal with them first.)

Self Assessment Questions

1) Fill in the blanks

   i) We would use _____, if we are testing a hypothesis of ì1 = ì2 and _______Test when ì1 = ì2= ì3 = ì4 if the populations under consideration are normally distributed.

   ii) ANOVA was developed by British Statistician __________________.

   iii) ANOVA is used when k __________.

   iv) ANOVA compares multiple means but the logic behind ANOVA is similar to _______ test that compares two independent means.

2) What are the assumptions of ANOVA?

   ....

....

....

....

3) Why are multiple *t*-tests not preferred when we have to compare more than 2 means?

....

....

....

....

3.3 INTRODUCTION TO KRUSKAL-WALLIS ANOVA TEST

When there are more than two groups or k number of groups to be compared, ANOVA is utilised but, again since ANOVA is a parametric statistics and requires assumption of normality as a key assumption, we need to also be aware of its non-parametric counterpart. The Kruskal-Wallis test compares the medians of several (more than two) populations to see whether they are all the same or not. The Kruskal Wallis test is a non-parametric analogue to ANOVA. It can be viewed as ANOVA based on rank transformed data.

That is, the initial data are transformed to their associated ranks before being subjected to ANOVA. In other words, it's like ANOVA, except that it is computed with medians and not means. It can also be viewed as a test of medians.

The null and alternative hypotheses may be stated as:

*H*0: the population medians are equal

*H*1: the population medians differ

3.4 RELEVANT BACKGROUND INFORMATION ON KRUSKAL WALLIS ANOVA TEST

The Kruskal-Wallis one way analysis of variance by ranks (Kruskal, 1952) and (Kruskal and Wallis, 1952) is employed with ordinal (rank order) data in a hypothesis testing situation involving a design with two or more independent samples. The test is an extension of the Mann-Whitney *U* test (Test 12) to a design involving more than two independent samples and, when k = 2, the Kruskal-Wallis one way analysis of variance by ranks will yield a result that is equivalent to that obtained with the Mann-Whitney *U* test.

If the result of the Kruskal-Wallis one-way analysis of variance by ranks is significant, it indicates there is a significant difference between at least two of the sample medians in the set of k medians. As a result of the latter, the researcher can conclude there is a high likelihood that at least two of the samples represent populations with different median values.

In employing the Kruskal-Wallis one-way analysis of variance by ranks one of the following is true with regard to the rank order data that are evaluated:

a) The data are in a rank-order format, since it is the only format in which scores are available; or

b) The data have been transformed into a rank-order format from an interval/ratio format, since the researcher has reason to believe that one or more of the assumptions of the single-factor between-subjects analysis of variance (which is the parametric analog of the Kruskal-Wallis test) are saliently violated.

It should be noted that when a researcher decides to transform a set of interval ratio data into ranks, information is sacrificed. This latter fact accounts for why there is reluctance among some researchers to employ non-parametric tests such as the Kruskal Wallis oneway analysis of variance by ranks, even if there is reason to believe that one or more of the assumptions of the single factor between subjects analysis of variance have been violated.

Various sources {e.g., Conover (1980, 1999), Daniel (1990), and Marascuilo and McSweeney (1977)} note that the Kruskal Wallis one way analysis of variance by ranks is based on the following assumptions:

a) Each sample has been randomly selected from the population it represents;

b) The k samples are independent of one another;

c) The dependent variable (which is subsequently ranked) is a continuous random variable. In truth, this assumption, which is common to many non-parametric tests, is often not adhered to, in that such tests are often employed with a dependent variable which represents a discrete random variable; and

d) The underlying distributions from which the samples are derived are identical in shape.

The shapes of the underlying population distributions, however, do not have to be normal.

Maxwell and Delaney (1990) point out that the assumption of identically shaped distributions implies equal dispersion of data within each distribution. Because of this, they note that, like the single factor between subjects analysis of variance, the Kruskal Wallis one way analysis of variance by ranks assumes homogeneity of variance with respect to the underlying population distribution. Because the latter assumption is not generally acknowledged for the Kruskal Wallis one way analysis of variance by ranks, it is not uncommon for sources to state that violation of the homogeneity of variance assumption justifies use of the Kruskal Wallis one way analysis of variance by ranks in lieu of the single factor between subjects analysis of variance.

It should be pointed out, however, that there is some empirical research which suggests that the sampling distribution for the Kruskal Wallis test statistic is not as affected by violation of the homogeneity of variance assumption as is the F distribution (which is the sampling distribution for the single-factor between-subjects analysis of variance).

One reason cited by various sources for employing the Kruskal Wallis one way analysis of variance by ranks is that by virtue of ranking interval/ratio data a researcher can reduce or eliminate the impact of outliers. As noted earlier in t test for two independent samples, since outliers can dramatically influence variability, they can be responsible for heterogeneity of variance between two or more samples. In addition, outliers can have a dramatic impact on the value of a sample mean.

Zimmerman and Zumbo (1993) note that the result obtained with the Kruskal-Wallis one-way analysis of variance by ranks is equivalent (in terms ofthe derived probability value) to that which will be obtained if the rank-orders employed for the Kruskal-Wallis test are evaluated with a single-factor between-subjects analysis of variance.

Self Assessment Questions

4) Fill in the balnks:

a) ANOVA is a parametric statistics its equivalent non-parametric statistics is ________________

b) Kruskal Wallis ANOVA was developed by ____________ and ____________ in 1952.

c) ANOVA compares means of more than two groups whereas ________ of more than two groups is compared by Kruskal Wallis ANOVA.

d) One of the assumptions in Kruskal Wallis AONA is that the dependent variable (which is subsequently ranked) is a ___________________

e) Kruskal Wallis ANOVA can be viewed as ANOVA based on _________ transformed data.

5) State the null and alternative hypothesis for Kruskal Wallis ANOVA.

....

....

....

....

6) Enumerate the assumptions of Kruskal Wallis ANOVA

....

....

....

....

3.5 STEP BY STEP PROCEDURE FOR KRUSKAL WALLIS ANOVA

1) Rank all the numbers in the entire data set from smallest to largest (using all samples combined); in the case of ties, use the average of the ranks that the values would have normally been given.

2) Total the ranks for each of the samples; call those totals *T*1, *T*2, . . ., *Tk*, where *k* is the number of groups or populations.

3) Calculate the Kruskal-Wallis test statistic,

$H = [\ 12 / N\ (N+1)\ ]\ [\ \Sigma((\Sigma R)^2 / n)\ ] - 3(N + 1)$

N = the total number of cases

n = the number of cases in a given group

$(\Sigma R)^2$ = the sum of the ranks squared for a given group of subjects

4) Find the *p*-value.

5) Make your conclusion about whether you can reject Ho by examining the *p*-value.

Example of a Small Sample:

In a Study, 12 participants were divided into three groups of 4 each, they were subjected to three different conditions, A (Low Noise), B(Avearge Noise), and C(Loud Noise). They were given a test and the errors committed by them on the test were noted and are given in the table below.

Participant No.	Condition A (Low Noise)	Participant No.	Condition B (Average Noise)	Participant No.	Condition C (Loud Noise)
1	3	5	2	9	10
2	5	6	7	10	8
3	6	7	9	11	7
4	3	8	8	12	11

The researcher wishes to know whether these three conditions differ amongst themselves. and there are no assumptions of the probability. To apply Kruskal Wallis test, following steps would be taken:

Step 1: Rank all the numbers in the entire data set from smallest to largest (using all samples combined); in the case of ties, use the average of the ranks that the values would have normally been given.

Condition A	Ranks T 1	Condition B	Ranks T2	Condition C	Ranks T3
3	2.5	2	1	10	11
5	4	7	6.5	8	8.5
6	5	9	10	7	6.5
3	2.5	8	8.5	11	12
	ΣT1 = 14		ΣT2 = 26		ΣT3 = 38

Step 2: Total the ranks for each of the samples; call those totals *T*1, *T*2, . . ., *Tk*, where *k* is the number of populations.

T1 =14

T2 =26

T3=38

Step3: Caculate H

$H = [\ 12 / N\ (N+1)\ ]\ [\ \Sigma((\Sigma R)^2 / n)\ ] – 3(N + 1)$

N = 12

n = 4

$(\Sigma R)^2 = (14+ 26+ 38)^2 = 6084$

H= [12/ 12 (12 + 1)] [(14^2/4) + (26^2/4) + (38^2/4)] – 3 (12+ 1)

H= [12/156] [49 + 169 + 361] – 39

H= (0.076 x 579) – 39

H= 44.525 – 39

H= 5.537

Step 4: Find the *p*-value.

Since the groups are three and number of items in each group are 4, therefore looking in table H (k=3, sample size of 4,4,4) it can be seen that the critical value is 5.692 (α = 0.05).

Step 5: Make your conclusion about whether you can reject Ho by examining the *p*-value.

Since the critical value is more than the actual value we *accept the null hypothesis* that all the three conditions A (Low Noise), B(Avearge Noise), and C(Loud Noise), do not differ from each other, therefore, in the said experiment there was no differences in the groups performance based on the noise level.

3.6 CONSIDERATIONS FOR LARGE SAMPLE

When the number of sample increases, the table H is unable to give us with the critical values, like for example it gives critical values up to 8 samples when k=3, 4 when k=4, and 3 samples when k=5, therefore as the sample increases table H is not of use for the critical value. In such a case we resort to chi square table for getting our information on the critical value taking degrees of freedom (k – 1).

Exact tables of the Kruskal-Wallis distribution: Although an exact probability value can be computed for obtaining a configuration of ranks which is equivalent to or more extreme than the configuration observed in the data evaluated with the Kruskal-Wallis one-way analysis of variance by ranks, the chi-square distribution is generally employed to estimate the latter probability. As the values of *k* and *N* increase, the chi-square distribution provides a more accurate estimate of the exact Kruskal-Wallis distribution. Although most sources employ the chi-square approximation regardless of the values of *k* and *N*, some sources recommend that exact tables be employed under certain conditions. Beyer (1968), Daniel, and Siegel and Castellan (1988) provide exact Kruskal-Wallis probabilities for whenever *k*= 3 and the number of subjects in any of the samples is five or less. Use of the chi-square distribution for small sample sizes will generally result in a slight decrease in the power of the test (i.e., there is a higher likelihood of retaining a false null hypothesis). Thus, for small sample sizes the tabled critical chi-square value should, in actuality, are a little lower than the value of Table H.

Worked Example for a large sample

A state court administrator asked the 24 court coordinators in the state’s three largest counties to rate their relative need for training in case flow management on a Likert scale (1 to 7).

1 = no training need

7 = critical training need

Training Need of Court Coordinators

County A	County B	County C
3	7	4
1	6	2
3	5	5
1	7	1
5	3	6
4	1	7
4	6	
2	4	
	4	
	5	

Step 1: Rank order the total groups' Likert scores from lowest to highest.

If tied scores are encountered, sum the tied positions and divide by the number of tied scores. Assign this rank to each of the tied scores.

Scores & Ranks Across the Three Counties

Ratings	Ranks	Ratings	Ranks
1	2.5	4	12
1	2.5	4	12
1	2.5	5	16.5
1	2.5	5	16.5
2	5.5	5	16.5
2	5.5	5	16.5
3	8	6	20
3	8	6	20
3	8	6	20
4	12	7	23
4	12	7	23
4	12	7	23

Calculating the ranks of tied scores

Example: Three court administrators rated their need for training as a 3. These three scores occupy the rank positions 7, 8, & 9.

$$(7 + 8 + 9) / 3 = 8$$

Step 2 Sum the ranks for each group and square the sums

County A		County B		County C	
Rating	Rank	Rating	Rank	Rating	Rank
3	8	7	23	4	12
1	2.5	6	20	2	5.5
3	8	5	16.5	5	16.5
1	2.5	7	23	1	2.5
5	16.5	3	8	6	20
4	12	1	2.5	7	23
4	12	6	20		
2	5.5	4	12		
		4	12		
		5	16.5		
Σ R $(\Sigma \text{ R})^2$	67.0 4489		153.5 23562.25		79.5 6320.25

Step 3 Calculate H

$H = [\ 12 / N\ (N+1)\ ]\ [\ \Sigma((\Sigma R)^2 / n)\ ] - 3(N + 1)$

$H = [\ 12 / 24\ (24+1)\ ]\ \ [4489 / 8 + 23562.25 / 10 + 6320.25 / 6] - 3\ (24 + 1)$

$H = (0.02)\ (3970.725) - (75)$

$H = 4.42$

$df = (k - 1) = (3 - 1) = 2$

Interpretation

The critical chi-square table value of H for $\alpha = 0.05$, and df = 2, is 5.991

Since 4.42 < 5.991, the null hypothesis is accepted. There is *no difference* in the training needs of the court coordinators in the three counties

Self Assessment Questions

1) Rearrange the following steps of Kruskal-Wallis test in appropriate order:

   i) Calculate H

   ii) Make your conclusion about whether you can reject Ho by examining the *p*-value.

   iii) Rank all the numbers in the entire data set from smallest to largest

   iv) Find the *p*-value.

   v) Total the ranks for each of the samples; call those totals *T*1, *T*2, . . ., *Tk*, where *k* is the number of populations.

2) Fill in the Blanks

i) As the values of *k* and *N* increase, the __________ distribution provides a more accurate estimate of the exact Kruskal-Wallis distribution.

ii) Use of the chi-square distribution for small sample sizes will generally result in a slight ____________ in the power of the test.

iii) When the critical value of H is more than the actual obtained value of H, we ____________ the null hypothesis.

iv) When the critical value of H is less than the actual obtained value of H, we ____________ the null hypothesis.

3.7 COMPARISON OF ANOVA AND KRUSKAL WALLIS ANOVA TEST

The Kruskal-Wallis (KW) ANOVA is the non-parametric equivalent of a one-way ANOVA. As it does not assume normality, the KW ANOVA tests the null hypothesis of no difference between three or more group medians, against the alternative hypothesis that a significant difference exists between the medians. The KW ANOVA is basically an extension of the Wilcoxon-Mann-Whitney (WMW) 2 sample test, and so has the same assumptions: 1) the groups have the same spreads; and 2) the data distributions have the same shape.

ANOVA compares means of different population to indicate the similarity between the populations KW ANOVA compares medians of these populations. ANOVA compares the data itself, KW ANOVA, converts data into ranks and then does its computation, in this respect Kruskal Wallis ANOVA is knows as Kruskal Wallis Analysis of Variance by Ranks.

Let’s look at the Example to see how their calculations differ:

Three groups 1, 2, and 3, performed a task, we want to see whether they differ or not.

Group 1	Group 2	Group 3		Group 1 (Ranks)	Group 2 (Ranks)	Group 3 (Ranks)
3	6	9		1	3.5	10
5	7	10		2	5.5	13
6	8	11		3.5	7.5	15.5
7	9	12		5.5	10	17
8	10	15		7.5	13	18
9	10			10	13	
	11				15.5	
38	61	57	Total	29.5	68	73.5

ANOVA

$n_1 = 6$ $n_2 = 7$ $n_3 = 5$

$\Sigma X_1 = 38$ $X_2 = 61$ $\Sigma X_3 = 57$

$\Sigma(X_1)^2 = 264$ $\Sigma (X_2)^2 = 551$ $\Sigma (X_3)^2 = 671$

$SS_{total} = (264 + 551 + 671) – [(38 + 61+ 57)^2 / 18] = 134$

$$SS_{\text{Between Groups}} = (38^2/6) + (61^2/7) + (57^2/5) – [(38 + 61+ 57)^2 / 18]$$

$$SS_{\text{Within Gropus}} = [\ 264 - (38^2/6)] + [\ 551- (61^2/7)]\ + [\ 671- (57^2/5)] = 63.962$$

Source of Variation	SS	df	MS	F ratio	F critical Value	Test Decision
Between Groups	70.038	2	35.019	8.213	3.68	Reject H_0
Within Groups	63.962	15	4.264			
Total	134.000	17				

Kruskal Wallis H test:

$H = [12 / 18(18+1)] [(29.5^2/6) + (68^2/7) + (73.5^2/5)] – [3 (18+1)]$

H= 66.177 – 57 = 9.177

Chi Square for Degrees of freedom 2 (3 – 1) is 5.99, Therefore reject the H_o

In both the case, ANOVA or Kruskal Wallis ANOVA, we will reject the Null Hypothesis, and state that the three groups differ.

F ratio as a function of H:

The fisher's F or F ratio or ANOVA one way variance is equivalent to H test or Kruskal Wallis test or Kruskal Wallis ANOVA, or ANOVA by rank order. This can also be seen in book Iman and Conover (1981)

Where the rank transform statistics states:

$F= [\{(k-1)/(N-k)\} \{((N-1)/H)-1\}]^{-1}$

If We see from the above mentioned example

F was 8.213 and H was 9.177

$F= [\{(3-1)/(18-3)\} \{((18-1)/9.177)-1\}]^{-1}$

$F= [(2/15) \{(17/9.177)-1\}]^{-1}$

$F= [0.133 \times (1.852-1)]^{-1} \quad = 0.1214^{-1}$

F= 8.231

3.8 LET US SUM UP

Kruskal Wallis One way Analysis of Variance (KW ANOVA) is a Non-parametric Analogue of ANOVA one way variance for independent sample. Kruskal Wallis ANOVA is used when there are more than 2 groups (k > 2). The assumption of KW ANOVA are that: a) Each sample has been randomly selected from the population it represents; b) The k samples are independent of one another; c) The dependent variable (which is subsequently ranked) is a continuous random variable. In truth, this assumption, which is common to many non-parametric tests, is often not adhered to, in that such tests are often employed with a dependent variable which represents a discrete random variable; and d) the underlying distributions from which the samples are derived are identical in shape.

The ANOVA or F test and KW ANOVA Or H test are equivalent to each other and can be appropriately used depending upon the type of population in question.

3.9 UNIT END QUESTIONS

1) Under what circumstances does the chi-square distribution provide an appropriate characterisation of the sampling distribution of the Kruskal–Wallis *H* statistic?

2) Data were collected from three populations—*A*, *B*, and *C*,—by means of a completely randomized design.

   The following describes the sample data:

   nA = nB = nC = 15

   RA = 235 RB = 439 RC = 361

   a) Specify the null and alternative hypotheses that should be used in conducting a test of hypothesis to determine whether the probability distributions of populations *A*, *B*, and *C* differ in location.

   b) Conduct the test of part a.

3) A firm wishes to compare four programs for training workers to perform a certain manual task. Twenty new employees are assigned to the training programs, with 5 in each program. At the end of the training period, a test is conducted to see how quickly trainees can perform the task. The number of times the task is performed per minute is recorded for each trainee, with the following results:

Observation	*Program 1*	*Program 2*	*Program 3*	*Program 4*
1	9	10	12	9
2	12	6	14	8
3	14	9	11	11
4	11	9	13	7
5	13	10	11	8

Calculate H, and report your results appropriately

4) An economist wants to test whether mean housing prices are the same regardless of which of 3 air-pollution levels typically prevails. A random sample of house purchases in 3 areas yields the price data below.

Mean Housing Prices (Thousands of Dollars):

MEAN HOUSING PRICES (THOUSANDS OF DOLLARS): *Pollution Level*			
Observation	Low	Mod	High
1	120	61	40
2	68	59	55
3	40	110	73
4	95	75	45
5	83	80	64

Calculate H and report your results with the p-value of 0.05

5) Show that H is equivalent to the F test statistics in one way analysis of variance problem if applied to the ranks of the observation rather than the actual numbers. (**Hint**: Express the F ratio as a function of H)

3.10 SUGGESTED READING AND REFERENCES

Daniel, W. W. (1990) *Applied Non-parametric Statistics*, 2d ed. Boston: PWS-Kent.

Iman, R. L., and W. J. Conover (1981), Rank transformations as a bridge between parametric and non-parametric statistics, The American Statistician, 35, 124–129.

Siegel S. and Castellan N.J. (1988) *Non-parametric Statistics for the Behavioral Sciences* (2nd edition). New York: McGraw Hill.

References

Johnson, Morrell, and Schick (1992), Two-Sample Non-parametric Estimation and Confidence Intervals Under Truncation, *Biometrics*, 48, 1043-1056.

Leach, C. (1979). Introduction to statistics: A non-parametric approach for the social sciences. Chichester: John Wiley & Sons

Lehman, E. L. (1975). Non-parametric statistical methods based on ranks. San Francisco: Holden-Day.

Wampold BE & Drew CJ. (1990) Theory and application of statistics. New York: McGraw-Hill.

UNIT 4 CHI-SQUARE AND KENDALL RANK CORRELATION

Structure

4.0 INTRODUCTION

In this unit, we will be discussing about the issues relating to the association and relationship between two or more variables. Generally when we want to measure the linear relationship between two variables, we apply Product Moment Coefficient of Correlation to the data and compute the ‘r’ value and check for its significance. This again we would do so if the data is normally distributed and the measurement of scores etc. are atleast in interval scale and there is a large sample. However if the sample size is small, and the distribution of the data is not known and the measurement is in nominal or ordinal scale, then we use non-parametric statistics related correlation, as for example the Rho or the Kendall Tau or where we need to know the association between two variables we may use the chi square test. In this unit we will be presenting first the measures of correlation both in parametric and non-parametric statistics, followed by Kendall rank order correlation, the Spearman Rank order correlation and the Chi Square test.

4.1 OBJECTIVES

On completing this unit, you will be able to:

- Define parametric and non-parametric tests of correlation;
- Explain the concepts underlying the non-parametric correlations;
- Describe the different non-parametric correlation techniques;
- Enumerate the step by step calculation of Kendall Tau; and
- Enumerate the step by step calculation of Chi Square test.

4.2 CONCEPT OF CORRELATION

The term “correlation” refers to a process for establishing whether or not relationships exist between two variables. Correlation quantifies the extent to which two quantitative variables, X and Y, “go together.” When high values of X are associated with high values of Y, a positive correlation exists. When high values of X are associated with low values of Y, a negative correlation exists. If values of X increases bringing about an increase in the values of Y simultaneously, X and Y are said to be positively correlated. If increases in X values bring about comparative decreases in the values of Y, then X and Y are said to be negatively correlated. If there is no typical trend in the increase or decrease of the variables then it is said to be not correlated or having zero correlation. Correlation ranges from -1 to 0 to +1. Correlation of +1.00 will indicate a perfect positive correlation and -1 will indicate a perfect negative correlation. Between these two extremes there could be many other degrees of correlation indicating positive or negative relationship between the variables. The correlation cannot exceed 1 in either direction. But it can have 0.54, 0.82, or 0.24, or 0.63 and so on at the positive level and at the negative level , it can have -0.55, -0.98, -0.67, -0.27 etc. All the latter are negative correlations and will not go beyond -1.00. Similarly the correlations that were mentioned as positive , will not exceed +1.00.

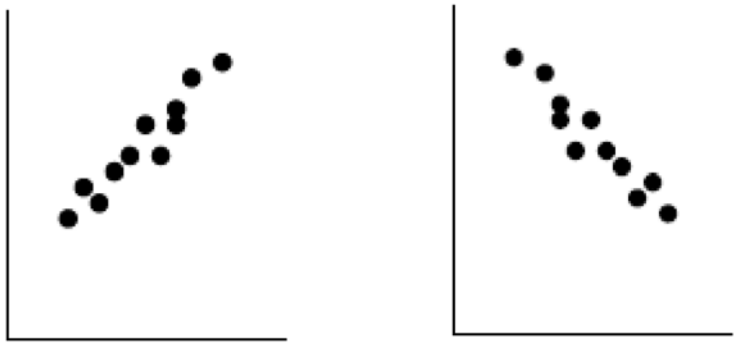
4.2.1 Scatter Plot

The first step is creating a scatter plot of the data. “There is no excuse for failing to plot and look.”

In general, scatter plots may reveal a

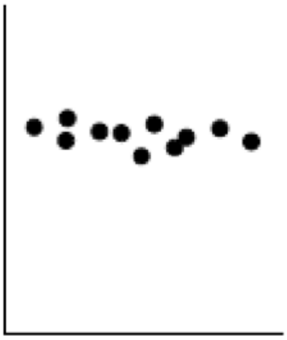
- positive correlation (high values of X associated with high values of Y)
- negative correlation (high values of X associated with low values of Y)
- no correlation (values of X are not at all predictive of values of Y).

These patterns are demonstrated in the figure below

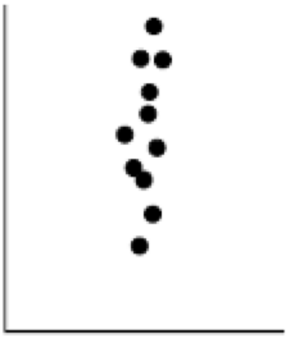


(A) Positive Correlation

(B) Negative Correlation



(A) No Correlation



(B) No Correlation

Correlation Coefficient

A single summary number that gives you a good idea about how closely one variable is related to another variable

This summary answers the following questions:

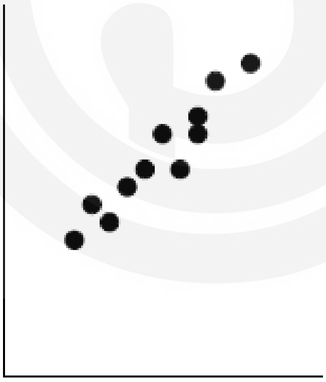
a) Does a relationship exist?

b) If so, is it a positive or a negative relationship? and

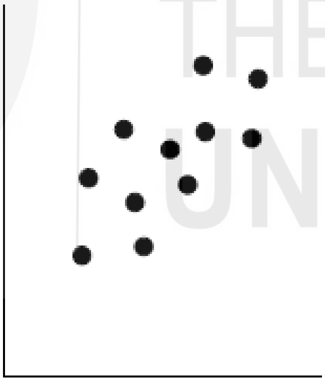
c) Is it a strong or a weak relationship?

Additionally, the same summary number would allow us to make accurate predictions about one variable when we have knowledge about the other variable.

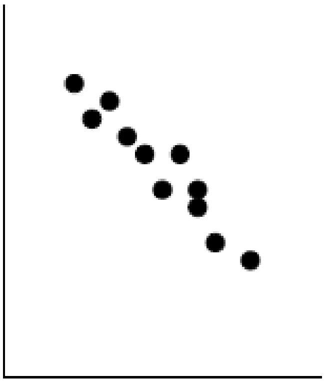
Correlation coefficients (denoted by r) are statistics that quantify the relationship between X and Y in unit free terms. When all points of a scatter plot fall directly on a line with an upward incline, $r = +1.00$, but when all points fall directly on a downward incline, $r =$



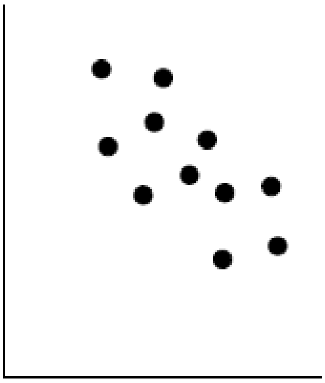
(A) Strong Positive Correlation



(B) Weak Positive Correlation



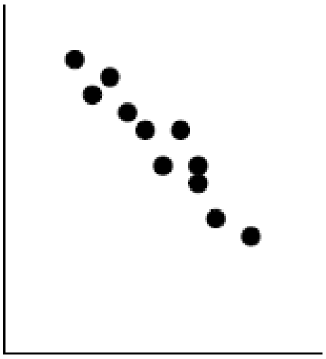
(C) Strong Negative Correlation



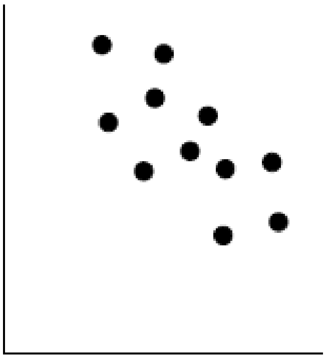
(D) Weak Negative Correlation

(A) Strong Positive Correlation

(B) Weak Positive Correlation



(A) Strong Negative Correlation



(B) Weak Negative Correlation

It is seen from the above that the strong correlation at both positive and negative directions is almost in a line with all the dots are placed very close to each other. On the other hand, the weak positive or negative correlation (refer to the graph above on the right hand side) that the points are placed far away from each other though the direction is somewhat clear. Thus there is a correlation but it appears rather weak.

4.2.2 Characteristics of Correlation

1) They tell you the direction of the relationship between two variables.

If your correlation coefficient is a negative number you can tell, just by looking at it, that there is a negative relationship between the two variables. As you may recall from the last chapter, a negative relationship means that as values on one variable increases (go up) the values on the other variable tend to decrease (go down) in a predictable manner.

If your correlation coefficient is a positive number, then you know that you have a positive relationship. This means that as one variable increases (or decreases) the values of the other variable tend to go in the same direction. If one increases, so does the other. If one decreases, so does the other in a predictable manner.

2) Correlation Coefficients always fall Between -1.00 and +1.00

All correlation coefficients range from -1.00 to +1.00. A correlation coefficient of -1.00 tells you that there is a *perfect negative relationship* between the two variables. This means that as values on one variable *increase* there is a perfectly predictable *decrease* in values on the other variable. In other words, as one variable goes up, the other goes in the opposite direction (it goes down).

A correlation coefficient of +1.00 tells you that there is a *perfect positive relationship* between the two variables. This means that as values on one variable *increase* there is a perfectly predictable *increase* in values on the other variable. In other words, as one variable goes up, so does the other.

A correlation coefficient of 0.00 tells you that there is a zero correlation, or no relationship, between the two variables. In other words, as one variable changes (goes up or down) you can't really say anything about what happens to the other variable. Sometimes the other variable goes up and sometimes it goes down. However, these changes are not predictable.

3) Larger Correlation Coefficients Mean Stronger Relationships

Most correlation coefficients (assuming there really is a relationship between the two variables you are examining) tend to be somewhat lower than plus or minus 1.00 (meaning that they are not perfect relationships) but are somewhat above 0.00. Remember that a correlation coefficient of 0.00 means that there is no relationship between the two variables based on the data given .

The closer a correlation coefficient is to 0.00, the weaker is the relationship and the less able one is to tell exactly what happens to one variable based on the knowledge of the other variable. The closer a correlation coefficient approaches plus or minus 1.00 the stronger the relationship is and the more accurately you are able to predict what happens to one variable based on the knowledge you have of the other variable.

4.3 MEASURES OF CORRELATION

4.3.1 Parametric Statistics

a) Pearson product moment correlation coefficient (Most widely accepted as a single appropriate statistics for correlation)

4.3.2 Non-parametric Statistics

a) Spearman's rank order correlation coefficient: Better known as "Spearman Rho" (Siegel & Castellan, 1988) assumes that the variables under consideration were measured on at least an ordinal (rank order) scale, that is, that the individual observations can be ranked into two ordered series. Spearman R can be thought of as the regular Pearson product moment correlation coefficient, that is, in terms of proportion of variability accounted for, except that Spearman R is computed from ranks.

b) Kendall's Tau: Explained in section 4.3.

c) Chi Square (Categorical Variables): Explained in section 4.7

Self Assessment Questions

1) Fill in the blanks:

i) Scatter plots may reveal a ___________correlation (high values of X associated with high values of Y)

ii) Scatter plots may reveal a __________correlation (high values of X associated with low values of Y)

iii) Scatter plots may reveal ________correlation (values of X are not at all predictive of values of Y).

iv) Correlation coefficients range from ______ to ______

v) A correlation coefficient of _____ tells you that there is a *perfect positive relationship* between the two variables.

vi) The closer a correlation coefficient is to 0.00, the __________ the relationship

vii) Correlation coefficient is a single summary number that gives you a good idea about *how closely one variable is* ___________ *to another variable*

2) What questions does correlation coefficient answers?

....

....

....

....

3) Name any two methods for calculating correlation?

....

....

....

....

4.4 KENDALL'S RANK ORDER CORRELATION (KENDALL'S TAU): (ð)

Kendall's tau (**ð**) is one of a number of measures of correlation or association. Measures of correlation are not inferential statistical tests, but are, instead, descriptive statistical measures which represent the degree of relationship between two or more variables. Upon computing a measure of correlation, it is a common practice to employ one or more inferential statistical tests in order to evaluate one or more hypotheses concerning the correlation coefficient. The hypothesis stated below is the most commonly evaluated hypothesis for Kendall's tau.

Null Hypothesis

H_o: **ð** $= 0$

(In the underlying population the sample represents, the correlation between the ranks of subjects on Variable X and Variable Y equals 0.)

4.4.1 Relevant Background Information on Test

Prior to reading the material in this section the reader should review the general discussion of correlation, of the Pearson product moment correlation coefficient and Spearman's rank order correlation coefficient (which also evaluates whether a monotonic relationship exists between two sets of ranks). Developed by Kendall (1938), tau is a bivariate measure of correlation/association that is employed with rank-order data. The population parameter estimated by the correlation coefficient will be represented by the notation **ð** (which is the lower case Greek letter tau). As is the case with Spearman's rank-order correlation coefficient Rho (***ñ***), Kendall's tau can be employed to evaluate data in which a researcher has scores for n subjects/objects on two variables (designated as the X and Y variables), both of which are rank-ordered.

Kendall's tau is also commonly employed to evaluate the degree of agreement between the rankings of m = 2 judges for n subjects/objects. As is the case with Spearman's rho, the range of possible values Kendall's tau can assume is defined by the limits - 1 to +1 (i.e., $-1 < r > +1$). Although Kendall's tau and Spearman's rho share certain properties in common with one another, they employ a different logic with respect to how they evaluate the degree of association between two variables.

Kendall's tau measures the degree of agreement between two sets of ranks with respect to the relative ordering of all possible pairs of subject/objects.

One set of ranks represents the ranks on the X variable, and the other set represents the ranks on the Y variable.

Specifically, The data are in the form of the following two pairs of observations expressed in a rank-order format:

a) $(R_x, R_y,)$ (which, respectively, represent the ranks on Variables X and Y for the 1st subject/object); and

b) $(R_{xj}, R_{yj},)$ (which, respectively, represent the ranks on Variables X and Y for the jth subject/object).

If the sign/direction of the difference $(R\ x_i - R\ _{yj})$, that is a pair of ranks is said to be concordant (i.e. in agreement).

If the sign/direction of the difference $(R_{xi} - R_{xj})$, a pair of ranks is said to be discordant (i.e., disagree).

If $(R_{yi} - R_{yj})$ and/or $(R_{xi} - R_{xj})$ result in the value of zero, a pair of ranks is neither the concordant nor discordant.

Kendall's tau is a proportion which represents the difference between the proportions of concordant pairs of ranks less the proportion of discordant pairs of ranks.

The computed value of tau will equal + 1 when there is complete agreement among the rankings (i.e., all of the pairs of ranks are concordant), and will equal -1 when there is complete disagreement among the rankings (i.e., all of the pairs of ranks are discordant).

As a result of the different logic involved in computing Kendall's tau and Spearman's rho, the two measures have different underlying scales, and, because of this, it is not possible to determine the exact value of one measure if the value of the other measure is known.

In spite of the differences between Kendall's tau and Spearman's rho, the two statistics employ the same amount of information, and, because of this, it is equally likely to detect a significant effect in a population.

In contrast to Kendall's tau, Spearman's rho is more commonly discussed in statistics books as a bivariate measure of correlation for ranked data. Two reasons for this are as follows:

a) The computations required for computing tau are more tedious than those required for computing rho; and

b) When a sample is derived from a bivariate normal distribution.

4.5 STEP BY STEP PROCEDURE FOR KENDALL RANK-ORDER CORRELATION

These are the steps in use of the Kendall rank order Correlation coefficient **ð(tau)**:

Rank the observations on the X variable from 1 to N. Rank the observations on the Y variable from 1 to N.

Arrange the list of N subjects so that the rank of the subjects on variable X are in their natural order, that is, 1, 2, 3,….N.

Observe the Y ranks in the order in which they occur when X ranks are in natural order. Determine the value of S, the number of agreements in order minus the number of disagreements in order, for the observed order of the Y ranks.

If there are no ties among either the X or the Y observations then we use the formula:

$T = 2S / (N (N -1))$

Where:

S = (score of agreement – score of disagreement on X and Y)

N = Number of objects or individuals ranked on both X and Y

If there are ties then the formula would be:

$T= 2S / [\quad N (N\text{-}1) – T_x \qquad T= 2S / [\quad N (N\text{-}1) – Ty$

Where:

S and N are as above

$T_x = \Sigma t (t – 1)$, t being the number of tied observations in each group of the ties on the X variable

$T_y = \Sigma t (t – 1)$, t being the number of tied observation in each group of the ties on the Y variable

If the N subjects constitute a random sample from some population, one may test the hypothesis that the variable X and Y are independent in that population. The method for doing so depends on the size of N:

For $N \leq 10$, Table — Upper tail probabilities for T, the Kendall rank order correlation coefficient

For N > 10, but less than 30, Table – Critical value for T, the Kendall rank order correlation coefficient

For N < 30 (or for intermediate significance levels for $10 < N \leq 30$) compute the value of z associated with T by using formula given below and use the z table

$z = 3T \quad N (N – 1) \; / \; 2 (2N+5)$

If the probability yielded by the appropriate method is equal to or less than the critical value, null hypothesis may be rejected in the favour of alternative hypothesis.

Worked up Example:

Without Ties:

Suppose we ask X and Y to rate their preference for four objects and give points out of 10. Now to see whether their preferences are related to each other we may use the following steps:

Data:

	A	B	C	D
X	6	8	5	2
Y	8	4	9	6

Step 1: Ranking the data of X and Y

	A	B	C	D
X	3	4	2	1
Y	3	1	4	2

Step 2: Rearrange the data of X in order of 1 to N (4 in this case)

	D	C	A	B
X	1	2	3	4

Step 3: Put the corresponding score of Y in order of X and Determine number of agreements and disagreements

	D	C	A	B
X	1	2	3	4
Y	2	4	3	1

To calculate S we need number of agreements and disagreements. This can be calculated by

Using the Y scores, starting from left and counting the number of ranks to its right that are larger, these are agreements in order. We subtract from this the number of ranks to its right that are smaller- these are the disagreements in order. If we do this for all the ranks and then sum the results we obtain S:

Y	2	4	3	1	Total
	2	+	+	-	+1
		4	-	-	-2
			3	-	-1
				1	0
				Grand Total= S	- 2

Step 4: Calculate T

T = 2S / (N (N -1))

T = 2 (– 2) / (4 (4 – 1))

T = – 4 / 12

T= – 0.33

Thus, T = – 0.33 is a measure of the agreement between the preferences of X and Y.

With Ties:

The two set of ranks to be correlated are:

Subject	A	B	C	D	E	F	G	H	I	J	K	L
Status striving rank	3	4	2	1	8	11	10	6	7	12	5	9
Yielding rank	1.5	1.5	3.5	3.5	5	6	7	8	9	10.5	10.5	12

As usual we would first rearrange X and observe the scores of corresponding Y scores to calculate S

Subject	D	C	A	B	K	H	I	E	L	G	F	J	
Status striving rank	1	2	3	4	5	6	7	8	9	10	11	12	
Yielding rank	3.5	3.5	1.5	1.5	10.5	8	9	5	12	7	6	10.5	Total
	3.5	0	-	-	+	+	+	+	+	+	+	+	8
		3.5	-	-	+	+	+	+	+	+	+	+	8
			1.5	0	+	+	+	+	+	+	+	+	8
				1.5	+	+	+	+	+	+	+	+	8
					10.5	-	-	-	+	-	-	0	-4
						8	+	-	+	-	-	+	0
							9	-	+	-	-	+	-1
								5	+	+	+	+	4
									12	-	-	-	-3
										7	-	+	0
											6	+	1
												10.5	0
											S=	Grand Total	25

We compute the value of S in usual way

$S = (8-2) + (8-2) + (8-0) + (8-0) + (1-5) +$

$(3-3) + (2-3) + (4-0) + (0-3) + (1-1) + (1-0) = 25$

It should be noted that, when there are tied observations, the ranks will be tied and neither rank in comparison pair precedes the other, so a value of 0 is assigned in the computation of S.

Having determined that S = 25, we now determine the value of T_x and T_Y. There are no ties among the scores on social status striving, i.e. in the X ranks and thus $T_x = 0$

On Y scores there are three sets of tied ranks. Two subjects are tied at 1.5, two subjects at 3.5, and two subjects' at 10.5 ranks. In each of these cases T = 2, the number of tied observations. Thus may be computed as:

$T_Y = \Sigma\, t\,(t - 1)$

$=2\,(2–1) + 2(2–1) + 2(2–1)$

$= 6$

With $T_x = 0$, $T_Y = 6$, $S = 25$, and $N = 12$, we may determine the value of T by using formula:

$T= 2S / [\sqrt{N\,(N - 1) - T_x}\ \sqrt{N\,(N - 1) - T_y}\ ]$

$T= (2 \times 25) / \sqrt{12\left(12 - 1\right) - 0}\ \sqrt{12\left(12 - 1\right) - 6}$

$= 0.39$

If we had not corrected the above coefficient for ties, i.e. we had used the previous formula for computing T we would have found T = 0.38. Observe that the effect of correcting for ties is relatively small unless the proportion of tied ranks is large or the number of ties in a group of ties is large.

4.6 FURTHER CONSIDERATIONS ABOUT KENDALL'S TAU

4.6.1 Comparison of Rho and Tau

For the example of tied observation if one calculates r it will be 0.62, whereas the T is 0.39. This example illustrates the fact that T and r have different underlying scales, and numerically they are not directly comparable to each other. That is if we measure the degree of correlation between A and B by using r and then do the same for A and C by using T, we cannot then say whether A is more closely related to B or to C because we have used noncomparable measures of correlation. It should be noted, however that there is a relation between the two measures which is best expressed in the following inequality:

$-1 \leq 3\,T - 2r \leq 1$

There are also differences in interpretation of the two measures. The spearman rank order correlation coefficient rho ***(ñ)*** is the same as a Pearson product moment correlation coefficient computed between variables the values of which consists of ranks. On the other hand, the Kendall rank-order correlation coefficient(**ð=tau)** has a different interpretation. It is the difference between the probability that, in the observed data, X and Y are in the same order and the probability that the X and Y data are in different orders. T_{XY} is different in the relative frequencies in the sample.

However, both coefficients utilise the same amount of information in the data, and thus both have the same sensitivity to detect the existence of association in the population. That is, the sampling distributions of T and r are such that for a given set of data both will lead to rejection of the null hypothesis at the same level of significance. However it should be remembered that the measures are different and measure association in different ways.

4.6.2 Efficiency of Rho

The Spearman Rho ***(ñ)*** and The Kendall (tau=ð) are similar in their ability to reject H_o, inasmuch as they make similar use of the information in the data.

When used on data to which the Pearson product moment correlation coefficient r is properly applicable, both Rho (***(ñ)*** and tau (**ð**) have efficiency of 91 percent. That is, Rho is approximately as sensitive a test of independence of two variables in a bivariate normal population with a sample of 100 cases as the Pearson r with 91 cases (Moran,1).

Self Assessment Questions

1) Fill in the blanks:

   i) Rho and tau have different underlying scales, and numerically they are not __________ to each other.

   ii) Developed by _____________ in year _________, tau is a __________measure of correlation/association that is employed with rank-order data.

2) State true or false:

   i) Kendall's tau measures the degree of agreement between two sets of ranks with respect to the relative ordering of all possible pairs of subject/objects.

   ii) Kendall's tau and Spearman's rho, the two measures have different underlying scales, and, because of this, it is not possible to determine the exact value of one measure if the value of the other measure is known.

   iii) Kendall's tau and Pearson's r both are rank order correlation, therefore both can be compared.

4.7 CHI-SQUARE TEST

The chi-square (X^2) test measures the alignment between two sets of frequency measures. These must be categorical counts and *not* percentages *or* ratios measures (for these, use another correlation test).

Note that the frequency numbers should be significant and be at least above 5 (although an occasional lower figure may be possible, as long as they are not a part of a pattern of low figures).

Chi Square performs two types of functions:

1) **Goodness of fit**

A common use is to assess whether a measured/observed set of measures follows an expected pattern. The expected frequency may be determined from prior knowledge (such as a previous year's exam results) or by calculation of an average from the given data.

The null hypothesis, H_0 is that the two sets of measures are not significantly different.

2) **Measure of Independence**

The chi-square test can be used in the reverse manner to goodness of fit. If the two sets of measures are compared, then just as you can show they align, you can also determine if they do *not* align.

The null hypothesis here is that the two sets of measures are similar.

The main difference in goodness-of-fit vs. independence assessments is in the use of the Chi Square table. For goodness of fit, attention is on 0.05, 0.01 or 0.001 figures. For independence, it is on 0.95 or 0.99 figures (this is why the table has two ends to it).

4.8 RELEVANT BACKGROUND INFORMATION ON TEST

The chi-square goodness-of-fit test, also referred to as the chi-square test for a single sample, is employed in a hypothesis testing situation involving a single sample. Based on some pre existing characteristic or measure of performance, each of *n* observations (subjects/objects) that is randomly selected from a population consisting of N observations (subjects/objects) is assigned to one of k mutually exclusive categories.' The data are summarized in the form of a table consisting of k cells, each cell representing one of the k categories.

The experimental hypothesis evaluated with the chi-square goodness-of-fit test is whether or not there is a difference between the observed frequencies of the k cells and their expected frequencies (also referred to as the theoretical frequencies). The expected frequency of a cell is determined through the use of probability theory or is based on some pre existing empirical information about the variable under study. If the result of the chi-square goodness-of-fit test is significant, the researcher can conclude that in the underlying population represented by the sample there is a high likelihood that the observed frequency for at least one of the k cells is not equal to the expected frequency of the cell. It should be noted that, in actuality, the test statistic for the chi-square goodness-of-fit test provides an approximation of a binomially distributed variable (when k = 2) and a multinomially distributed variable (when k > 2). The larger the value of *n*, the more accurate the chi-square approximation of the binomial and multinomial distributions.

The chi-square goodness-of-fit test is based on the following assumptions:

a) Categorical nominal data are employed in the analysis. This assumption reflects the fact that the test data should represent frequencies for k mutually exclusive categories;

b) The data that are evaluated consists of a random sample of n independent observations. This assumption reflects the fact that each observation can only be represented once in the data; and

c) The expected frequency of each cell is 5 or greater.

When this assumption is violated, it is recommended that if $k = 2$, the binomial sign test for a single sample be employed to evaluate the data. When the expected frequency of one or more cells is less than 5 and $k > 2$, the multinomial distribution should be employed to evaluate the data. The reader should be aware of the fact that sources are not in agreement with respect to the minimum acceptable value for an expected frequency.

Many sources employ criteria suggested by Cochran (1952), who stated that none of the expected frequencies should be less than 1 and that no more than 20% of the expected frequencies should be less than 5. However, many sources suggest the latter criteria may be overly conservative. In the event that a researcher believes that one or more expected cell frequencies are too small, two or more cells can be combined with one another to increase the values of the expected frequencies.

Zar (1999) provides an interesting discussion on the issue of the lowest acceptable value for an expected frequency. Within the framework of his discussion, Zar (1999) cites studies indicating that when the chi-square goodness-of-fit test is employed to evaluate a hypothesis regarding a uniform distribution, the test is extremely robust.

A robust test is one that still provides reliable information, in spite of the fact that one or more of its assumptions have been violated. A uniform distribution (also referred to as a rectangular distribution) is one in which each of the possible values a variable can assume has an equal likelihood of occurring. In the case of an analysis involving the chi-square goodness-of-fit test, a distribution is uniform if each of the cells has the same expected frequency.

4.9 STEP BY STEP PROCEDURE FOR CHI-SQUARE TEST

1) Write the observed frequencies in column *O*

2) Figure the expected frequencies and write them in column *E*.

Expected Frequencies:

When you find the value for chi square, you determine whether the observed frequencies differ significantly from the expected frequencies. You find the expected frequencies for chi square in three ways:

1) You hypothesize that all the frequencies are equal in each category. For example, you might expect that half of the entering freshmen class of 200 at Tech College will be identified as women and half as men. You figure the expected frequency by dividing the number in the sample by the number of categories. In this exam pie, where there are 200 entering freshmen and two categories, male and female, you divide your sample of 200 by 2, the number of categories, to get 100 (expected frequencies) in each category.

2) You determine the expected frequencies on the basis of some prior knowledge. Let us use the Tech College example again, but this time pretend we have prior knowledge of the frequencies of men and women in each category from last year's entering class, when 60% of the freshmen were men and 40% were women. This year you might expect that 60% of the total would be men and 40% would be women. You find the expected frequencies by multiplying the sample size by each of the hypothesized population proportions. If the freshmen total were 200, you would expect 120 to be men (60% x 200) and 80 to be women (40% x 200).

3) Use the formula to find the chi-square value:

   Chi Square = $\Sigma\ [(O - E)^2 / E]$

   Where:

   O is the Observed Frequency in each category

   E is the Expected Frequency in the corresponding category

4) Find the *df*. (*N*-1)

5) Find the table value (consult the Chi Square Table.)

6) If your chi-square value is *equal to or greater than* the table value, reject the null hypothesis: *differences in your data are not due to chance alone*

Worked Up Example:

Situation: Mr. X., the manager of a car dealership, did not want to stock cars that were bought less frequently because of their unpopular color. The five colors that he ordered were red, yellow, green, blue, and white. According to Mr. X, the expected frequencies or number of customers choosing each color should follow the percentages of last year. She felt 20% would choose yellow, 30% would choose red, 10% would choose green, 10% would choose blue, and 30% would choose white. She now took a random sample of 150 customers and asked them their colour preferences. The results of this poll are shown in Table below under the column labelled as observed frequencies."

Category Color	Observed Frequencies	Expected Frequencies
Yellow	35	30
Red	50	45
Green	30	15
Blue	10	15
White	25	45

The expected frequencies in Table are figured from last year's percentages. Based on the percentages for last year, we would expect 20% to choose yellow. Figure the expected frequencies for yellow by taking 20% of the 150 customers, getting an expected frequency of 30 people for this category. For the colour red we would expect 30% out of 150 or 45 people to fall in this category.

Using this method, Thai figured out the expected frequencies 30, 45, *15*, *15*, and 45. Obviously, there are discrepancies between the colours preferred by customers in the poll taken by Mr.X. and the colours preferred by the customers who bought their cars *last* year. Most striking is the difference in the green and white colours. If Thai were to follow the results of her poll, she would stock twice as many green cars than if she were to follow the customer colour preference for green based on last year's sales. In the case of white cars, she would stock half as many this year. What to do? Mr. X. needs to know whether or not the discrepancies between last year's choices (expected frequencies) and this year's preferences on the basis of his poll (observed frequencies) demonstrate a *real* change in customer colour preferences. It could be that the differences are simply a result of the random sample she *chanced to* select. If so, then the population of customers really has not changed from last year as far as colour preferences go.

The *null hypothesis* states that there is no significant difference between the expected and observed frequencies.

The *alternative hypothesis* states they *are* different. The level of significance (the point at which you can say with 95% confidence that the difference is NOT due to chance alone) is set at .05 (the standard for most science experiments.) The chi-square formula used on these data is

Chi Square = $\Sigma\ [(O - E)^2 / E]$

Where:

O is the Observed Frequency in each category

E is the Expected Frequency in the corresponding category

df is the "degree of freedom" (n-1)

We are now ready to use our formula for X^2 and find out if there *is* a significant difference between the observed and expected frequencies for the customers in choosing cars. We will set up a worksheet; then you will follow the directions to form the columns and solve the formula.

1) *Directions for Setting up Worksheet for Chi Square*

Category	O	E	O-E	$(O-E)^2$	$(O-E)^2$ / E
Yellow	35	30	5	25	0.83
Red	50	45	5	25	0.56
Green	30	15	15	225	15
Blue	10	15	-5	25	1.67
White	25	45	-20	400	8.89
				Total=	26.95

This Total is the Chi Square value. After calculating the Chi Square value, find the *"Degrees of Freedom."*

(Remember: DO *NOT* SQUARE THE NUMBER YOU GET, NOR FIND THE SQUARE ROOT - THE NUMBER YOU GET FROM COMPLETING THE CALCULATIONS AS ABOVE IS CHI SQUARE.)

2) *Degrees of freedom (df)* refers to the number of values that are free to vary after restriction has been placed on the data. For instance, if you have four numbers with the restriction that their sum has to be 50, then three of these numbers can be anything, they are free to vary, but the fourth number *definitely* is restricted. For example, the first three numbers could be 15, 20, and 5, adding up to 40; then the fourth number has to be 10 in order that they sum to 50. The degrees of freedom for these values are then three. The degrees of freedom here is defined as *N* - 1, the number in the group minus one restriction (4 - 1).

3) Find the table value for Chi Square. Begin by finding the *df* found in step 2 along the left hand side of the table. Run your fingers across the proper row until you reach the predetermined level of significance (.05) at the column heading on the top of the table. The table value for Chi Square in the correct box of *4 df* and *P=.05* level of significance is 9.49.

4) If the calculated chi-square value for the set of data you are analysing (26.95) is equal to or greater than the table value (9.49), reject the null hypothesis. *There is a significant difference between the data sets that cannot be due to chance alone.* If the number you calculate is LESS than the number you find on the table, then you can probably say that any differences are due to chance alone.

In this situation, the rejection of the null hypothesis means that the differences between the expected frequencies (based upon last year's car sales) and the observed frequencies (based upon this year's poll taken by Mr.X) are not due to chance. That is, they are not due to chance variation in the sample Mr.X took. There is a real difference between them. Therefore, in deciding what colour autos to stock, it would be to Mr.X's advantage to pay careful attention to the results of her poll!

Another Example:

Let us take an example of Males and Females in two different categories, full stop and rolling stop and no stop. Now to see whether they are different from each other or more similar to each other we will follow the following steps

Step 1: Add numbers across columns and rows. Calculate total number in chart.

Unobtrusive Male Versus Female

	Male	Female	
Full Stop	6	6	= 12
Rolling Stop	16	15	= 31
No Stop	4	3	= 7
	= 26	= 24	= 50

Step 2: Calculate the expected numbers for each individual cell. Do this by multiplying row sum by column sum and dividing by total number. For example: using 1st cell in table (Male/Full Stop);

12 x 26 / 50 = 6.24

2nd cell in table (Female/Full Stop):

12 x 24 / 50 = 5.76

Step 3: Now you should have an observed number and expected number for each cell. The observed number is the number already in 1st chart. The expected number is the number found in the last step (step 2). Sometimes writing both numbers in the chart can be helpful

	Male	Female	
Full Stop	6 (observed) 6.24 (expected)	6 (observed) 5.76 (expected)	= 12
Rolling Stop	16 (observed) 16.12 (expected)	15 (observed) 14.88 (expected)	= 31
No Stop	4 (observed) 3.64 (expected)	3 (observed) 3.36 (expected)	= 7
	= 26	= 24	= 50

Step 4:

Chi Square = Sum of (Observed - Expected)^2 / Expected

Calculate this formula for each cell, one at a time. For example, cell #1 (Male/Full Stop):

Observed number is: 6 Expected number is: 6.24

Plugging this into the formula, you have:

$(6 – 6.24)^2 /6.24 = .0092$

Continue doing this for the rest of the cells, and add the final numbers for each cell together for the final Chi Square number. There are 6 total cells, so at the end you should be adding six numbers together for you final Chi Square number.

Step 5: Calculate degrees of freedom (*df*):

(Number of Rows – 1) x (Number of Columns – 1)

(3 – 1) x (2 – 1)

2 x 1 =

2 *df* (degrees of freedom)

Step 6: Look up the number in the chart at end of handout. At .05 significance level, with 2 *df*, the number in chart should be 5.99. Therefore, in order to reject the null hypothesis, the final answer to the Chi Square must be greater or equal to 5.99. The Chi Square/final answer found was .0952. This number is less than 5.99, so you fail to reject the null hypothesis, thus there is no difference in these groups.

4.10 FURTHER CONSIDERATIONS ABOUT CHI SQUARE

Observations must appear in one cell only. For instance, if we looked at male and female swimmers and hurdlers, one person could appear in both the swimmers *and* the hurdlers category if they enjoyed both sports. This would make use of Chi square invalid. Actual frequencies must appear in the cells, not percentages, proportions or numbers which do anything other than count. For instance, the mean of an interval scale variable cannot appear.

LOW expected frequencies

One limitation is that one should not proceed with a chi square test where expected frequency cells fall below 5. The rule of thumb which most statisticians inherited, and which comes from Cochran (1954) which was that *no more than 20% of expected cells should fall below* 5. This would rule out any 2 X 2 in which at least one expected cell was less than 5.

Hypothetical table:

Age	**Conversed**	**Did not converse**	**Total**
5 years	2	6	8
7 years	6	2	8
Total	8	8	16

For total sample sizes less than 20 and two expected cells below *5*, the risk of a type I error is too high. For instance, the data shown in hypothetical table above give a *chi square* of 4.0 (which is 'significant' for one *df)* yet it's easy to see, again, without much formal statistical training, that the result was relatively likely to occur - only two children in each age group needed to move away, in opposite directions, from the expected frequencies of four in each cell for these results to occur. From first principles (working out all the possible combinations) the probability of these results occurring comes out substantially higher than 0.05. If you have these sort of data it doesn't take too long to work from first principles but it's far better to make sure your analysis will be valid by taking a large enough sample, with a sensible design. Even with tables larger than 2X2, if several expected frequencies fall below 5 and the row or column total are quite severely skewed, the possibility of a type I error increases.

Self Assessment Questions

1) What are the assumptions of chi-square goodness-of-fit test?

....

....

....

....

2) Chi square performs two major functions, what are these?

....

....

....

....

3) State true or false:

i) The expected frequency of a cell is determined through the use of probability theory or is based on some pre existing empirical information about the variable under study.

ii) If several expected frequencies fall below 5, the possibility of a type II error increases.

iii) The chi-square (c^2) test measures the alignment between two sets of frequency measures.

iv) “The data that are evaluated consists of a random sample of n independent observations.” Is not a cardinal assumptions of chi square?

4.11 LET US SUM UP

In this unit we learnt about the concept of correlation and how parametric test is used to compute the product moment coefficient of correlation. We thedn learnt about the non-parametric tests for corrleation and leatrnt about the Rho and Tau. The Rho was by Spearman and was known as Spearman Rank Correlation while Kendall’s Tau was known as ð (tau). We also learnt about how to calculate Kendall’s tau and learnt about the importance of Chi-Square test. We also learnt as to how to calculate chi-square.

4.12 UNIT END QUESTIONS

1) Compute correlation coefficient for each of the following pairs of sample observations:

a)

x	33	61	20	19	40
y	26	36	65	25	35

b)

x	89	102	120	137	41
y	81	94	75	52	136

c)

x	2	15	4	10
y	11	2	15	21

d)

x	5	20	15	10	3
y	80	83	91	82	87

2) Compare T and r in terms of correlation and state your views?

3) Should a chi-Square test be carried out on the following data?

7	1
2	7

4) A (fictitious) Survey shows that. in a sample of 100.9 I people are against the privatisation of health services, whereas 9 support the idea.

   a) What test of significance can be performed on this data?

   b) Calculate the chi square value and check it for significance.

   c) Could this test be one-tailed?

   If for a large sample, we knew *on/y* that 87% of people were against the idea and were for could we carry out the same test to see whether this split is significant

5) What is the difference between chi square goodness of fit test and measure of independence test?

6) What do you understand by efficiency of T?

4.13 SUGGESTED READINGS

Daniel, W. W. (1990) *Applied Non-parametric Statistics*, 2d ed. Boston: PWS-Kent.

Johnson, Morrell, and Schick (1992), Two-Sample Non-parametric Estimation and Confidence Intervals Under Truncation, *Biometrics*, 48, 1043-1056.

Siegel S. and Castellan N.J. (1988) *Non-parametric Statistics for the Behavioral Sciences* (2nd edition). New York: McGraw Hill.

Wampold BE & Drew CJ. (1990) *Theory and Application of Statistics*. New York: McGraw-Hill.