NON-PARAMETRIC STATISTICS *

(http://www.Statsoft.com)

1. GENERAL PURPOSE

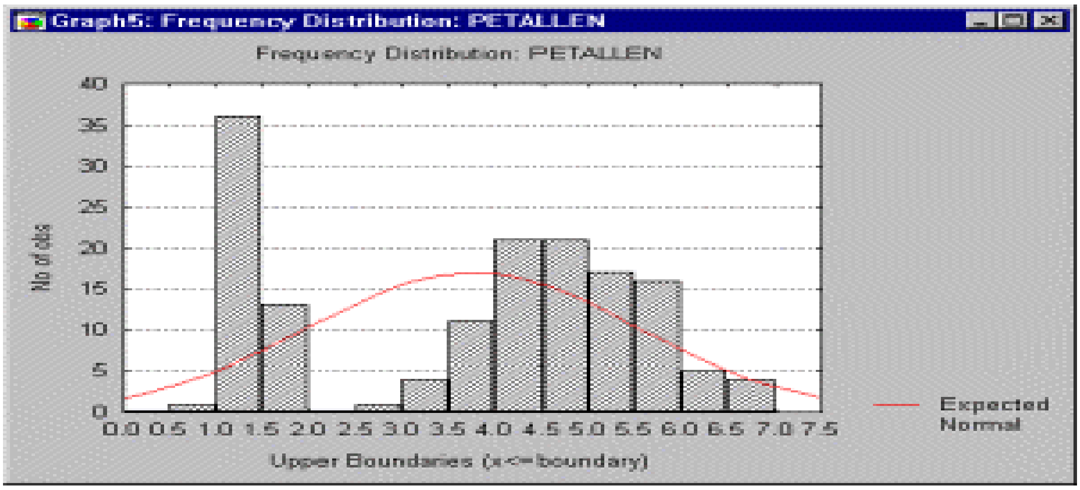
1.1 Brief review of the idea of significance testing

To understand the idea of non-parametric statistics (the term non-parametric was first used by Wolfowitz, 1942) first requires a basic understanding of parametric statistics. The Elementary Concepts chapter of the manual introduces the concept of statistical significance testing based on the sampling distribution of a particular statistic (you may want to review that chapter before reading on). In short, if we have a basic knowledge of the underlying distribution of a variable, then we can make predictions about how, in repeated samples of equal size, this particular statistic will "behave," that is, how it is distributed. For example, if we draw 100 random samples of 100 adults each from the general population, and compute the mean height in each sample, then the distribution of the standardized means across samples will likely approximate the normal distribution (to be precise, Student's t distribution with 99 degrees of freedom; see below). Now imagine that we take an additional sample in a particular city ("Tallburg") where we suspect that people are taller than the average population. If the mean height in that sample falls outside the upper 95% tail area of the t distribution then we conclude that, indeed, the people of Tallburg are taller than the average population.

1.2 Are most variables normally distributed?

In the above example we relied on our knowledge that, in repeated samples of equal size, the standardized means (for height) will be distributed following the t distribution (with a particular mean and variance). However, this will only be true if in the population the variable of interest (height in our example) is normally distributed, that is, if the distribution of people of particular heights follows the normal distribution (the bell-shape distribution).

*** This material is a verbatim presentation.**



For many variables of interest, we simply do not know for sure that this is the case. For example, is income distributed normally in the population? – Probably not. The incidence rates of rare diseases are not normally distributed in the population, the number of car accidents is also not normally distributed, and neither are very many other variables in which a researcher might be interested.

For more information on the normal distribution, see Elementary Concepts; for information on tests of normality, see Normality tests.

1.3 Sample size

Another factor that often limits the applicability of tests based on the assumption that the sampling distribution is normal is the size of the sample of data available for the analysis (sample size; n). We can assume that the sampling distribution is normal even if we are not sure that the distribution of the variable in the population is normal, as long as our sample is large enough (e.g., 100 or more observations). However, if our sample is very small, then those tests can be used only if we are sure that the variable is normally distributed, and there is no way to test this assumption if the sample is small.

1.4 Problems in measurement

Applications of tests that are based on the normality assumptions are further limited by a lack of precise measurement. For example, let us consider a study where grade point average (GPA) is measured as the major variable of interest. Is an A average twice as good as a C average? Is the difference between a B and an A average comparable to the difference between a D and a C average? Somehow, the GPA is a crude measure of scholastic accomplishments that only allows us to establish a rank ordering of students from "good" students to "poor" students. This general measurement issue is usually discussed in statistics textbooks in terms of types of measurement or scale of measurement. Without going into too much detail, most common statistical techniques such as analysis of variance (and t-tests), regression, etc. assume that the underlying measurements are at least of interval, meaning that equally spaced intervals on the scale can be compared in a meaningful manner (e.g, B minus A is equal to D minus C). However, as in our example, this assumption is very often not tenable, and the data rather represent a rank ordering of observations (ordinal) rather than precise measurements.

1.5 Parametric and non-parametric methods

Hopefully, after this somewhat lengthy introduction, the need is evident for statistical procedures that allow us to process data of "low quality," from small samples, on variables about which nothing is known (concerning their distribution). Specifically, non-parametric methods were developed to be used in cases when the researcher knows nothing about the parameters of the variable of interest in the population (hence the name non-parametric). In more technical terms, non-parametric methods do not rely on the estimation of parameters (such as the mean or the standard deviation) describing the distribution of the variable of interest in the population. Therefore, these methods are also sometimes (and more appropriately) called parameter-free methods or distribution-free methods.

2. BRIEF OVERVIEW OF NON-PARAMETRIC METHODS

Basically, there is at least one non-parametric equivalent for each parametric general type of test. 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

2.1 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 in Basic Statistics. Non-parametric alternatives for this test are the Wald-Wolfowitz run 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 Kruskal-Wallis analysis of ranks and the Median test.

2.2 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 (in Basic Statistics for example, if we wanted to compare students' math skills at the beginning of the semester with their skills at the end of the semester). Non-parametric alternatives to this test are the Sign test and **Wilcoxon's matched pair 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.

2.3 Relationships between variables

To express a relationship between two variables one usually computes the correlation coefficient. Non-parametric equivalents to the standard correlation coefficient are Spearman R, Kendall Tau, and coefficient Gamma (see Non-parametric correlations). If the two variables of interest are categorical in nature (e.g., "passed" vs. "failed" by "male" vs. "female") appropriate non-parametric statistics for testing the relationship between the 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: Kendall coefficient of concordance. This test is often used for expressing inter-rater agreement among independent judges who are rating (ranking) the same stimuli.

2.4 Descriptive statistics

When one's data are not normally distributed, and the measurements at best contain rank order information, then computing the standard descriptive statistics (e.g., mean, standard deviation) is sometimes not the most informative way to summarize the data. 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.

3. WHEN TO USE WHICH METHOD

It is not easy to give simple advice concerning the use of non-parametric procedures. Each non-parametric procedure has its peculiar sensitivities and blind spots. For example, the Kolmogorov-Smirnov two-sample test is not only sensitive to differences in the location of distributions (for example, differences in means) but is also greatly affected by differences in their shapes. The Wilcoxon matched pairs test assumes that one can rank order the magnitude of differences in matched observations in a meaningful manner. If this is not the case, one should rather use the Sign test. In general, if the result of a study is important (e.g., does a very expensive and painful drug therapy help people get better?), then it is always advisable to run different non-parametric tests; should discrepancies in the results occur contingent upon which test is used, one should try to understand why some tests give different results. On the other hand, non-parametric statistics are less statistically powerful (sensitive) than their parametric counterparts, and if it is important to detect even small effects (e.g., is this food additive harmful to people?) one should be very careful in the choice of a test statistic.

3.1 Large data sets and non-parametric methods

Non-parametric methods are most appropriate when the sample sizes are small. When the data set is large (e.g., n > 100) it often makes little sense to use non-parametric statistics at all. The Elementary Concepts chapter of the manual briefly discusses the idea of the central limit theorem. In a nutshell, when the samples become very large, then the sample means will follow the normal distribution even if the respective variable is not normally distributed in the population, or is not measured very well. Thus, parametric methods, which are usually much more sensitive (i.e., have more statistical power) are in most cases appropriate for large samples. However, the tests of significance of many of the non-parametric statistics described here are based on asymptotic (large sample) theory; therefore, meaningful tests can often not be performed if the sample sizes become too small. Please refer to the descriptions of the specific tests to learn more about their power and efficiency.

4. NON-PARAMETRIC CORRELATIONS

The following are three types of commonly used non-parametric correlation coefficients (Spearman R, Kendall Tau, and Gamma coefficients). Note that the chi-square statistic computed for two-way frequency tables, also provides a careful measure of a relation between the two (tabulated) variables, and unlike the correlation measures listed below, it can be used for variables that are measured on a simple nominal scale.

- **Spearman R**

  Spearman R (Siegel & Castellan, 1988) assumes that the variables under consideration were measured on atleast an ordinal (rank order) scale, that is, that the individual observations can be ranked into two ordered series. Spearman Rcan 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.

- **Kendall tau**

  Kendall tau is equivalent to Spearman R with regard to the underlying assumptions. It is also comparable in terms of its statistical power. However, Spearman R and Kendall tau are usually not identical in magnitude because their underlying logic as well as their computational formulas are very different. Siegel and Castellan (1988) express the relationship of the two measures in terms of the inequality:

  $-1 \leq 3 * \text{Kendall tau} - 2 * \text{Spearman R} \leq 1$

  More importantly, Kendall tau and Spearman R imply different interpretations: 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. Kendall tau, on the other hand, represents a probability, that is, it is the difference between the probability that in the observed data the two variables are in the same order versus the probability that the two variables are in different orders.

- **Gamma**

The Gamma statistic (Siegel & Castellan, 1988) is preferable to Spearman R or Kendall tau when the data contain many tied observations. In terms of the underlying assumptions, Gamma is equivalent to Spearman R or Kendall tau; in terms of its interpretation and computation it is more similar to Kendall tau than Spearman R. In short, Gamma is also a probability; specifically, it is computed as the difference between the probability that the rank ordering of the two variables agree minus the probability that they disagree, divided by 1 minus the probability of ties. Thus, Gamma is basically equivalent to Kendall tau, except that ties are explicitly taken into account.