7 NON-PARAMETRIC STATISTICS

7.1 ANDERSON - DARLING TEST: The **Anderson–Darling test** is a statistical test of whether a given sample of data is drawn from a given probability distribution. In its basic form, the test assumes that there are no parameters to be estimated in the distribution being tested, in which case the test and its set of critical values is distribution-free. However, the test is most often used in contexts where a family of distributions is being tested, in which case the parameters of that family need to be estimated and account must be taken of this in adjusting either the test-statistic or its critical values. When applied to testing if a normal distribution adequately describes a set of data, it is one of the most powerful statistical tools for detecting most departures from normality. ***K*-sample Anderson–Darling tests** are available for testing whether several collections of observations can be modeled as coming from a single population, where the distribution function does not have to be specified.

In addition to its use as a test of fit for distributions, it can be used in parameter estimation as the basis for a form of minimum distance estimation procedure.

The test is named after Theodore Wilbur Anderson (born 1918) and Donald A. Darling (born 1915), who invented it in 1952.

The Anderson-Darling test for normality is one of three general normality tests designed to detect all departures from normality. While it is sometimes touted as the most powerful test, no one test is best against all alternatives and the other 2 tests are of comparable power. The p-values given by Distribution Analyzer for this test may differ slightly from those given in other software packages as they have been corrected to be accurate to 3 significant digits.

The test rejects the hypothesis of normality when the p-value is less than or equal to 0.05. Failing the normality test allows you to state with 95% confidence the data does not fit the normal distribution. Passing the normality test only allows you to state no significant departure from normality was found.

The Anderson-Darling test, while having excellent theoretical properties, has a serious flaw when applied to real world data. The Anderson-Darling test is severely affected by ties in the data due to poor precision. When a significant number of ties exist, the Anderson-Darling will frequently reject the data as non-normal, regardless of how well the data fits the normal distribution. Below is an example of data generated from the normal distribution but rounded to the nearest 0.5 to create ties. A tie is when identical values occurs more than once in the data set.

7.2 COHEN'S KAPPA COEFFICIENT: Cohen's kappa coefficient is a statistical measure of inter-rater agreement or *inter-annotator agreement* for qualitative (categorical) items. It is generally thought to be a more robust measure than simple percent agreement calculation since κ takes into account the agreement occurring by chance. Some researchers have expressed concern over κ's tendency to take the observed categories' frequencies as givens, which can have the effect of underestimating agreement for a category that is also commonly used; for this reason, κ is considered an overly conservative measure of agreement.

Others contest the assertion that kappa "takes into account" chance agreement. To do this effectively would require an explicit model of how chance affects rater decisions. The so-called chance adjustment of kappa statistics supposes that, when not completely certain, raters simply guess—a very unrealistic scenario.

A case sometimes considered to be a problem with Cohen's Kappa occurs when comparing the Kappa calculated for two pairs of raters with the two raters in each pair having the same percentage agreement but one pair give a similar number of ratings while the other pair give a very different number of ratings.

STATISTICAL SIGNIFICANCE makes no claim on how important is the magnitude in a given application or what is considered as high or low agreement.

Statistical significance for kappa is rarely reported, probably because even relatively low values of kappa can nonetheless be significantly different from zero but not of sufficient magnitude to satisfy investigators. Still, its standard error has been described and is computed by various computer programs.

If statistical significance is not a useful guide, what magnitude of kappa reflects adequate agreement? Guidelines would be helpful, but factors other than agreement can influence its magnitude, which makes interpretation of a given magnitude problematic. As Sim and Wright noted, two important factors are prevalence (are the codes equiprobable or do their probabilities vary) and bias (are the marginal probabilities for the two observers similar or different). Other things being equal, kappas are higher when codes are equiprobable. On the other hand Kappas are higher when codes are distributed asymmetrically by the two observers. In contrast to probability variations, the effect of bias is greater when Kappa is small than when it is large.

Another factor is the number of codes. As number of codes increases, kappas become higher. Based on a simulation study, Bakeman and colleagues concluded that for fallible observers, values for kappa were lower when codes were fewer. And, in agreement with Sim & Wrights's statement concerning prevalence, kappas were higher when codes were roughly equiprobable. Thus Bakeman et al. concluded that "no one value of kappa can be regarded as universally acceptable."They also provide a computer program that lets users compute values for kappa specifying number of codes, their probability, and observer accuracy. For example, given equiprobable codes and observers who are 85% accurate, value of kappa are 0.49, 0.60, 0.66, and 0.69 when number of codes is 2, 3, 5, and 10, respectively.

Nonetheless, magnitude guidelines have appeared in the literature. Perhaps the first was Landis and Koch, who characterized values < 0 as indicating no agreement and 0–0.20 as slight, 0.21–0.40 as fair, 0.41–0.60 as moderate, 0.61–0.80 as substantial, and 0.81–1 as almost perfect agreement. This set of guidelines is however by no means universally accepted; Landis and Koch supplied no evidence to support it, basing it instead on personal opinion. It has been noted that these guidelines may be more harmful than helpful. Fleiss's equally arbitrary guidelines characterize kappas over 0.75 as excellent, 0.40 to 0.75 as fair to good, and below 0.40 as poor.

7.3. TEST OF FRIEDMAN: The **Friedman test** is a non-parametric statistical test developed by the U.S. economist Milton Friedman. Similar to the parametric repeated measures ANOVA, it is used to detect differences in treatments across multiple test attempts. The procedure involves ranking each row (or *block*) together, then considering the values of ranks by columns. Applicable to complete block designs, it is thus a special case of the Durbin test.

Classic examples of use are:

- *n* wine judges each rate *k* different wines. Are any wines ranked consistently higher or lower than the others?
- *n* wines are each rated by *k* different judges. Are the judges' ratings consistent with each other?
- *n* welders each use *k* welding torches, and the ensuing welds were rated on quality. Do any of the torches produce consistently better or worse welds?

The Friedman test is used for one-way repeated measures analysis of variance by ranks. In its use of ranks it is similar to the Kruskal-Wallis one-way analysis of variance by ranks.

Friedman test is widely supported by many statistical software packages.

7.4. TEST OF KOLMOGOROV – SMIRNOV: In statistics, the **Kolmogorov–Smirnov test (K–S test)** is a nonparametric test of the equality of continuous, one-dimensional probability distributions that can be used to compare a sample with a reference probability distribution (one-sample K–S test), or to compare two samples (two-sample K–S test). The Kolmogorov–Smirnov statistic quantifies a distance between the empirical distribution function of the sample and the cumulative distribution function of the reference distribution, or between the empirical distribution functions of two samples. The null distribution of this statistic is calculated under the null hypothesis that the samples are drawn from the same distribution (in the two-sample case) or that the sample is drawn from the reference distribution (in the one-sample case). In each case, the distributions considered under the null hypothesis are continuous distributions but are otherwise unrestricted.

The two-sample K–S test is one of the most useful and general nonparametric methods for comparing two samples, as it is sensitive to differences in both location and shape of the empirical cumulative distribution functions of the two samples.

The Kolmogorov–Smirnov test can be modified to serve as a goodness of fit test. In the special case of testing for normality of the distribution, samples are standardized and compared with a standard normal distribution. This is equivalent to setting the mean and variance of the reference distribution equal to the sample estimates, and it is known that using these to define the specific reference distribution changes the null distribution of the test statistic: see below. Various studies have found that, even in this corrected form, the test is less powerful for testing normality than the Shapiro–Wilk test or Anderson–Darling test. However, other tests have their own disadvantages. For instance the Shapiro-Wilk test is known not to work well with many ties (many identical values)

7.5.TEST OF KRUSKAL – WALLIS: The **Kruskal–Wallis one-way analysis of variance** by ranks (named after William Kruskal and W. Allen Wallis) is a non-parametric method for testing whether samples originate from the same distribution. It is used for comparing more than two samples that are independent, or not related. The parametric equivalent of the Kruskal-Wallis test is the one-way analysis of variance(ANOVA). When the Kruskal-Wallis test leads to significant results, then at least one of the samples is different from the other samples. The test does not identify where the differences occur or how many differences actually occur. It is an extension of the Mann–Whitney U test to 3 or more groups. The

Mann-Whitney would help analyze the specific sample pairs for significant differences.

Since it is a non-parametric method, the Kruskal–Wallis test does not assume a normal distribution of the residuals, unlike the analogous one-way analysis of variance. However, the test does assume an identically shaped and scaled distribution for each group, except for any difference in medians.

7.6. TEST OF MANN-WHITNEY: n statistics, the **Mann–Whitney *U* test** (also called the **Mann–Whitney–Wilcoxon** (**MWW**), **Wilcoxon rank-sum test**, or **Wilcoxon–Mann–Whitney test**) is a nonparametric test of thenull hypothesis that two populations are the same against an alternative hypothesis, especially that a particular population tends to have larger values than the other.

It has greater efficiency than the t-test on non-normal distributions, such as a mixture of normal distributions, and it is nearly as efficient as the t-test on normal distributions.

The Wilcoxon rank-sum test is not the same as the Wilcoxon signed-rank test, although both are nonparametric and involve summation of ranks.

7.7. THE MEDIAN TEST: In statistics, Mood's **median test** is a special case of Pearson's chi-squared test. It is a nonparametric test that tests the null hypothesis that the medians of the populations from which two or more samples are drawn are identical. The data in each sample are assigned to two groups, one consisting of data whose values are higher than the median value in the two groups combined, and the other consisting of data whose values are at the median or below. A Pearson's chi-squared test is then used to determine whether the observed frequencies in each sample differ from expected frequencies derived from a distribution combining the two groups.

7.8 SPEARMAN'S RANK CORRELATION COEFFICIENT:
In statistics, Spearman's rank correlation coefficient or Spearman's rho, named after Charles Spearman and often denoted by the Greek letter ρ (rho) or as r_s, is a nonparametric measure of statistical dependence between two variables. It assesses how well the relationship between two variables can be described using a monotonic function. If there are no repeated data values, a perfect Spearman correlation of +1 or −1 occurs when each of the variables is a perfect monotone function of the other.

Spearman's coefficient, like any correlation calculation, is appropriate for both continuous and discrete variables, including ordinalvariables

7.9 WILCOXON SIGNED RANKS TEST: The **Wilcoxon signed-rank test** is a non-parametric statistical hypothesis test used when comparing two related samples, matched samples, or repeated measurements on a single sample to assess whether their population mean ranks differ (i.e. it is a paired difference test). It can be used as an alternative to the paired Student's t-test, *t*-test for matched pairs, or the *t*-test for dependent samples when the population cannot be assumed to be normally distributed.

The Wilcoxon signed-rank test is not the same as the Wilcoxon rank-sum test, although both are nonparametric and involve summation of ranks.

The test is named for Frank Wilcoxon (1892–1965) who, in a single paper, proposed both it and the rank-sum test for two independent samples (Wilcoxon, 1945). The test was popularized by Siegel (1956) in his influential text book on non-parametric statistics. Siegel used the symbol *T* for a value related to, but not the same as, W. In consequence, the test is sometimes referred to as the **Wilcoxon *T* test**, and the test statistic is reported as a value of *T*.

Assumptions

1. Data are paired and come from the same population.
2. Each pair is chosen randomly and independently.
3. The data are measured at least on an ordinal scale, but need not be normal.