A R O N I S O F T   L L C

AroniSmartLytics™ Handbook of Applied Statistics

P.O. Box 1104
Fair Lawn, New Jersey 07410, United States of America
E-Mail: aronisoft@aroni.us   Web: http://www.aroni.us

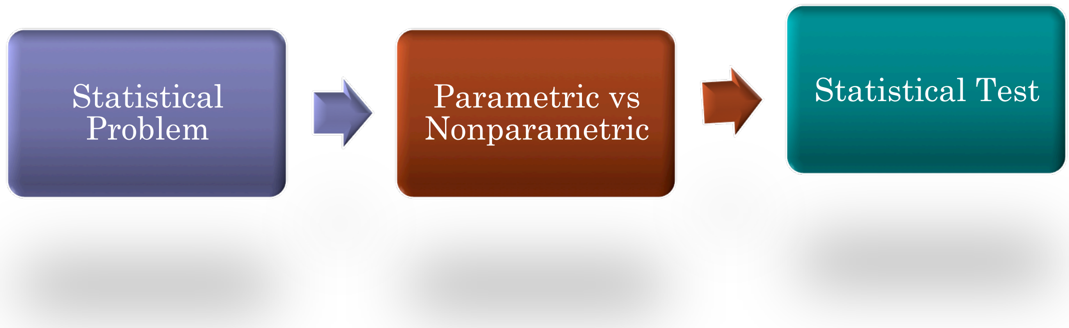
I. The Statistics World With *AroniSmartLytics*™

Researchers, Scientists, Engineers, Students, Politicians, Economists, Political Analysts, Financial Analysts, Statisticians, and most of us the practitioners use statistics to answer questions that require data analysis. Statistics help to describe the results of, among others, investigations, experiments, observations. When data about the subject or object of interest is somehow comprehensive, data analysis using ***descriptive statistics*** is what is needed. Unfortunately, in the world of statistics, things are not that simple. The population of interest may not be fully observable or accessible. Hence, the statistician will have to make inference, or in statistics terms, appeal to ***inferential statistics***. Inferential statistics involve statistics tests. One of the major issues facing statistics researchers is to choose which test to use. The choice involves first a decision between two families of test statistics: ***parametric*** and ***nonparametric***. AroniSmartLytics™ intends to help beginners, practitioners and most advanced statisticians and researchers to navigate this tricky step in learning, research, and interpretation and application of or action on results.

The specific module of AroniSmartLytics™ dedicated to statistic tests offers an in-depth, user friendly and intuitive reference and selection tool that will facilitate the task, save time, while being rigorously thought out.

- Statistics vs. Statistical Tests
- Non-Parametric Statistical Tests
- Parametric Statistical Tests
- Non-parametric vs Parametric Statistical Tests: when to choose one versus the other

The module follows the simple three steps:



II. Statistics

The statistical method usually follows the five steps of the scientific method: **State the problem, Formulate hypothesis, Design experiment or survey, Make observations, Interpret data, Draw conclusions**. Statistics covers the aspects of the collection, organization, analysis, and interpretation of data. These aspects include the planning of data collection in terms of the design of surveys, researches, and experiments. In order to analyze data, hypotheses, and findings, statisticians use **statistics**, which are the quantity calculated or assumed from the data, such as: **mean, proportion, median, variance, standard deviation, moment**.

To make inference, statisticians test hypotheses or conjectures, using specific **statistical tests** on given "**statistics**". These tests are usually grouped into two categories: *non-parametric tests* and ***parametric tests***, hence Nonparametric and Parametric statistics.

III. Nonparametric Statistics

Nonparametric statistics deals with techniques to analyze data that do not belong to any particular probability distribution. The list of key probability distributions is given in the main module of ***AroniSmartLytics™.***

Data that do not belong to any distribution can be analyzed using distribution free methods, which do not rely on any assumptions regarding a probability distribution that can describe the data. These distribution free methods are covered in nonparametric statistics. *AroniSmartLytics™* will give the most common and uncommon methods or statistical tests used in nonparametric statistics.

In ***nonparametric statistics***, the structure of the model, and the number and nature of the parameters are flexible and not fixed in advance. They are determined from the data instead of being specified a priori.

In ***parametric statistics***, distribution parameters are assumed in advance.

IV. Parametric Statistics

Unlike *nonparametic* statistics, ***parametric statistics*** assumes that data have come from a type of probability distribution, and only makes inference about the parameters of the distribution. The list of key probability distributions is given in the main module of ***AroniSmartLytics™.***

Methods based on parametric statistics make more assumptions than nonparametric statistics and produce estimates under specific assumptions. If the assumptions are correct, the parametric statistics accurately describes the research problem. Hence, parametric statistics commonly lacks or has limited robusteness. Usually, parametric statistics uses formulas to describe the probability distributions, their parameters, and the statistics, making their nature simple.

V. Nonparametric vs Parametric Statistics.

The choice between ***Nonparametric*** and ***parametric statistics*** may be tricky and non obvious. ***AroniSmartLytics™*** simplifies the problem. The choice of a statistics test is based on mostly three criteria:

- ***Nature of the problem or goal***
- ***Levels or Scale of Measurement and nature of the data***
- ***Robustness***

The **combination** of the *nature of the problem* and the *scale of measurement* determine the family of statistical tests to use. Robustness determines the type of the test, and ultimately the test itself.

A. Nature of the problem.

Statistical problems may be grouped into the following nine categories.

- Describing one group
- Comparing one group to a hypothetical value
- Comparing two unpaired groups
- Comparing two paired groups

- Comparing three or more unmatched groups
- Comparing three or more matched groups
- Quantifying association between two variables
- Predicting value from another measured variable
- Predicting value from several measured or binomial variables

The most basic independent group design has two groups. These are often called ***Experimental*** and ***Control*** group. In two-group settings, subjects or objects are randomly selected from the population and randomly assigned each to one or the other of two groups. Sometimes, the subjects from two different groups are paired. Other times, there is no basis for pairing scores, and the groups are independent. Two independent groups may or may not have the same number of subjects.

B. Levels or Scale of Measurements or Types of Data

There are five levels of measurements: **Binomial or Two possible outcomes, Nominal or discrete unordered, Ordinal** or **Discrete ranked, Numerical discrete, Numerical continuous or ratio**, and **survival**

C. Robusteness.

Given the nature of a problem, some tests are more robust that others. ***AroniSmartLytics™*** will help narrow down the range to the smallest set of possible tests. Faced with multiple choices, by using ***AroniSmartLytics™*** the researcher, statistician, student, engineer and practitioner will have it easier when selecting the most appropriate and relevant test based on a rigorous and tested approach.

VI. Choosing between Nonparametric and Parametric tests.

The first question that may come into mind when trying to decide between the two types of statistical tests, that is Nonparametric and Parametric, is whether the choice is critical. The short answer is that it matters mostly in the following situations:

A. Small samples

- Using nonparametric tests for small samples that are suited for parametric tests (Gaussian populations) may lead to diminished robusteness, hence reduced sample power to make inference.
- Using parametric tests with small samples that are not suited (non-Gaussian populations) may lead to inaccurate inference, given that the central limit theorem is no longer applicable. Probability statements obtained from most nonparametric statistics are usually exact probabilities, regardless of the shape of the population distribution from which the random sample was drawn.
- If too small sample sizes are used, there is no alternative to using a nonparametric test.

B. Large Samples

- Usually with large samples, it is easier. Parametric tests with large samples from non-Gaussian populations may be protected by the central limit theorem. The problem is that the concept of “large sample” may be a matter of taste, with 30 subjects or more being commonly used a cutoff between small and large samples.
- Corollary to this, nonparametric tests on Gaussian populations usually deteriorate the robustness of the test.

AroniSmartLytics™ has a clear, user-friendly module that helps to select which parametric test to use and in what situations nonparametric statistical tests may be appropriate. The module clarifies the steps involved in choosing between a nonparametric and parametric test.

Once a decision has been made to use nonparametric tests, the nonparametric statistics module will guide in selecting the appropriate nonparametric test.

Similarly, a function within the software will guide the user in selecting the most suitable parametric statistical test.

In any case, by using ***AroniSmartLytics™*** and ***AroniStat™,*** researchers, scientists, engineers, students, politicians, economists, political analysts, financial Analysts, statisticians, and most practitioners may select and examine the shapes and forms of probability density and cumulative density functions and probability mass and cumulative mass functions. By varying the parameter values of the probability distributions, they may observe various shapes and forms that will help them select the distribution of interest. This will help in selecting the parametric tests and consequently the appropriate distribution to use in parametric tests.

VII. Main Nonparametric Tests.

A. Binomial Test

The ***Binomial Tests*** are appropriate for populations consisting of only two classes. Many populations have these characteristics: *male* and *female*, *married* and *single*, *member* and *nonmember*, *inpatient* and *outpatient*, *literate* and *illiterate*, etc. Such populations are called ***binary or dichotomous populations***.

Statistic/ Concept	Formula	Comment
Test/ Variable Definition	$P(X = x\|p) =$ $p^x(1-p)^{1-x}$; $x = 0, 1;\ \ 0 \leq p \leq 1$	$Pr(x = 1)$ $= 1 - \Pr(X = 0)$ $= 1 - q$ $= p$
Expected Value	p	$0 \leq p \leq 1$
Variance	$p(1-p)$	pq
Common Hypothesis	$Ho{:}p = {}^1/_2$	Symmetry of Binomial when p=1/2: $P(X \geq x) = P(X \leq n-x)$
Small samples	$Ho{:}p = {}^1/_2$	
Large Samples	$(X - n_p)\Big/\sqrt{npq}$	$pq = p(1-q)$
Asymptotic Distribution	$Z = \dfrac{(X - N_p)}{\sqrt{npq}}$	Z is approximately normally distributed with mean 0 and standard deviation 1
Correction for Continuity	$Z = \frac{(X \pm 0.5) - Np}{\sqrt{npq}}$ When X<Np use X+.5 and when X<Np use X-.5	
Power	Best test for dichotomous data	

B. Binomial Confidence Intervals

Binomial Confidence Intervals are used in assessing the confidence interval for a proportion in a population with two distinct groups. The proportion statistic is estimated from a sample of observations. The sampling error is calculated based on the observations using a formula. Several formulae exist. For all of them, the assumption is that the population proportion follows a Binomial Distribution.

The methods to compute Binomial Confidence Intervals include:

- *Normal approximation interval*
- *Wilson score interval*
- *Clopper-Pearson interval*
- *Agresti-Coull Interval*
- *Jeffreys interval*

1. Normal approximation interval

Normal approximation intervals are the most commonly used method. The method assumes that the a binomial distribution may be approximated by a normal distribution, as shown in the module of Relationships Among Key Probability Distributions of ***AroniSmartLytics™.***

The approximation relies on the Central limit theorem. Other assumptions are that both the proportions of success or failure should not be close to zero.

Statistic/ Concept	Formula	Comment
Test/ Concept Definition	$\hat{p} \pm z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ $z_{1-\frac{\alpha}{2}}$ *is the* $1-\frac{\alpha}{2}$ *percentile of a standard normal distribution*	$\hat{p}$ *is sample proportion of success in a Bernouilli trial* α *is the error perecentile* n *is the sample size*
Hypothesis Testing and Test Validity	*The population proportion* π, *is tested as follows:* $\left\{\pi \middle| z_{\alpha/2} \le \frac{\hat{p}-\pi}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \le z_{1-\alpha/2}\right\}$	
Other Assumptions and	$\frac{\hat{p}-\pi}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}$	Is know as Wald test, hence the test is also know as Wald Test

2. Wilson score interval

The Wilson interval, named after its developer, the American Mathematical Statistician Edwin Bidwell Wilson, was proposed as an improvement to the Normal approximation. The test is widely viewed as closer to the true population value. The assumptions remain the same as those for the Normal approximation.

Statistic/ Concept	Formula	Comment
Test/ Concept Definition	$\frac{\hat{p} \pm \frac{1}{2n} z^2{}_{1-\alpha/2} \pm z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{z^2{}_{1-\alpha/2}}{4n^2}}}{1 + \frac{1}{2n} z^2{}_{1-\alpha/2}}$ $z_{1-\frac{\alpha}{2}}$ *is the* $1-\frac{\alpha}{2}$ *percentile of a standard normal distribution* $\hat{p}$ *is sample proportion of success in a Bernouilli trial* α *is the error perecentile* n *is the sample size*	
Hypothesis Testing and Test Validity	*The population proportion* π, *is tested as follows:* $\left\{ \pi \middle\vert z_{\alpha/2} \le \frac{\hat{p}-\pi}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \le z_{1-\alpha/2} \right\}$	
Other Assumptions and	$\frac{\hat{p} \pm \frac{1}{2n} z^2{}_{1-\alpha/2}}{1 + \frac{1}{2n} z^2{}_{1-\alpha/2}}$	is known as the Wilson test

3. Clopper-Pearson interval

The Clopper-Pearson interval, named after the British statisticians Egon Sharpe Pearson and C. Clopper, is also known as "**Exact Confidence Interval**." Like the Wilson Test, the Clopper-Pearson interval is proposed as an improvement to the Normal Approximation. The method is based on the cumulative probabilities of the binomial distribution.

The method approximates the intervals by two binomial cumulative mass functions: a **lower bound** and **an upper bound**.

The interval may be estimated by the **Beta Inverse distribution** and the **F distribution**. The relationship between the **Beta and F distributions** is shown in the module of **Relationships Among Key Probability Distributions** of *AroniSmartLytics™.*

Statistic/ Concept	Formula	Comment
Test/ Concept Definition	$\{\pi \mid P[Binomial(n;\pi) \leq X] \geq \alpha/2\} \cap \{\pi \mid P[Binomial(n;\pi) \geq X] \geq \alpha/2\}$	
Hypothesis Testing and Test Validity	$The\ population\ proportion\ \pi, is\ tested\ as\ follows:$ $\left\{\sum_{k=0}^{k}\binom{n}{k}p_{UB}^{k}(1-p_{UB})^{n-\square}=\alpha/2\right\}\cap\left\{\sum_{k=x}^{n}\binom{n}{k}p_{LB}^{k}(1-p_{LB})^{n-k}=\alpha/2\right\}$ The population proportion, π, falls in the range $[p_{LB;}p_{UB}]$ where: p_{LB} is the confidence interval lower bound p_{UB} is the confidence interval upper bound n is the number of trials or sample size k is the number of successes in n trials	
Other Assumptions and	The confidence intervals may be estimated by the Beta Distribution: $p_{UB}=1-BetaInverse\left(\frac{1-\alpha}{2},n-k,k+1\right)$ $p_{LB}=1-BetaInverse\left(1-\frac{1-\alpha}{2},n-k+1,k\right)$	

4. Agresti-Coull Interval

The Agresti-Coull Interval Approximation, named after the American mathematical statisticians Alan Agresti and Brent A. Coull, is the approximate binomial confidence interval that is also an improvement on the Normal approximation. It is closely related to the Wilson estimation.

Statistic/ Concept	Formula	Comment
Test/ Concept Definition	$\tilde{p} \pm z_{1-\alpha/2}\sqrt{\frac{\tilde{p}(1-\tilde{p})}{\tilde{n}}}$; $z_{1-\frac{\alpha}{2}}$ *is the* $1-\frac{\alpha}{2}$ *percentile of a standard normal distribution* $\tilde{p} = \frac{X + z^2_{1-\frac{\alpha}{2}}\overline{2}}{\tilde{n}}$ *and* $\tilde{n} = n + z^2_{1-\frac{\alpha}{2}}$, *where n is the sample size* *X is the number of successes in n trials*	
Hypothesis Testing and Test Validity	*The population proportion* π, *is tested as follows*: $\left\{\pi \middle\vert z_{\alpha/2} \le \frac{\tilde{p}-\pi}{\sqrt{\frac{\tilde{p}(1-\tilde{p})}{\tilde{n}}}} \le z_{1-\alpha/2}\right\}$	
Other Assumptions and		

5. Jeffreys interval

The Jeffreys interval, named after the British statistician mathematician, geophysicist and astronomer Harold Jeffreys is based on Bayesian statistics. It is derived from the Jeffreys prior binomial proportion p, which is non informative objective or empirical prior distribution parameter space that is proportional to the square root of the determinant of the Fisher information.

For the Binomial proportion p, the Jeffreys prior is a Beta distribution with parameters (1/2, 1/2).

The estimation process of the Jeffreys interval follows the steps:

- **Step 1:** observe x successes in n trials;

- **Step 2:** Form the posterior probability distribution for p as a Beta distribution with parameters $(x + 1/2, n - x + 1/2)$;
- **Step 3:** For $x \neq 0$ and $x \neq n$, the Jeffreys interval is estimated by the $100(1 - \alpha)\%$ equal-tailed posterior probability interval, that is the $\alpha / 2$ and $1 - \alpha / 2$ quantiles of a Beta distribution with parameters $(x + 1/2, n - x + 1/2)$.

Estimation Considerations: Quantiles may be computed using ***AroniSmartLytics*** probability distributions module.

To avoid estimations degradations, especially the coverage probability tending to zero when $p \to 0$ or 1, when $x = 0$, the upper limit is calculated whereas the lower limit is set to 0, and when $x = n$ the lower limit is calculated, whereas the upper limit is set to 1.

C. Binomial Power Tests

Binomial Power test plays a major role in assessing the strength in hypothesis testing. When testing hypothesis, there are two possible decisions: rejecting or accepting **the null hypothesis**.

The decision to reject the hypotheses after observing the data and calculating the statistic may not be the final step in the decision making process. In fact, the null hypothesis is rejected when the observed statistic falls beyond a critical value. However, even after the decision is made further confirmation may be needed to strengthen the confidence in the decision: **is the sample size enough or is more or stronger evidence needed.**

Hence the statistical problem may be represented by the following table:

Decision Based on the Observations	True Situation	
	Ho: True	Ho: False
Fail to Reject the null hypothesis Ho	Correct	Type II Error ß
Reject the null hypothesis Ho	Type I error α	Correct

In practice, the two types of errors are not given the same importance or weight. The decision is biased against rejecting a true null hypothesis. The null hypothesis should not be rejected without strong evidence.

The researcher wants the Probability (Type I Error) = Probability (Rejecting Ho when Ho is true) to be very small. Fortunately the Probability (type I error), denoted by α, can be determined once the form of the critical region is known.

At the same time, the researcher wants to avoid making the Type II error, which the probability of deciding not to reject Ho when Ho is false.

The Type II error, commonly denoted by ß, is a function of the proportion of successes p, the sample size n, and the Type I error α. Hence, the aim is to maximize the chances of not making the Type II error. This chance is known as the Power of a Test

Power of a Test

Let p be the test statistic and Å, the rejection region for a test of a hypothesis for the population (true) parameter π. Then the power of the statistic test is the probability that the test will lead to rejection of Ho when the actual parameter value is π. Let W be the statistical test.

That is, power (π) = Probability (W in Å when the parameter value is π) or Probability of rejecting the Ho when Ho is false.

Power =1-probabilty of Type II error = $\mathbf{1-\beta}$

D. Chi-Squared Goodness-of-Fit Test

The ***Chi-Squared (or Chi-square) Goodness-of-Fit*** *test* is suited for populations consisting of various categories when researchers, students and practitioners are interested in the number of subjects, objects, or responses that fall into the various categories. The categories may be two or more. The rationale of the Chi-squared Goodness-of-Fit test is to test whether there exist significant differences between ***observed*** and ***expected*** number of objects, subjects, or responses falling into each category.

Statistic/ Concept	Formula	Comment
Test/ Variable Definition	$X^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} = \sum_{i=1}^{k} \frac{{O_i}^2}{E_i} - n$ $k \geq 2\,; k < n$ n= number of observations	O_i= observed number of cases for category i E_i= expected number of cases for category I when Ho is true k = total number of categories
Expected Value and Variance	See chi-square distribution	
Common Hypothesis	$Ho: X^2\ follows\ chi - square\ (\chi^2)$ $distribution\ with\ degrees$ $of\ freedom\ df = k - 1$	When Ho assumes equal proportion of cases in each category $E_i = n/k$ df is adjusted for estimated parameters
Small Expected Frequencies	Usually expect frequency is at least 5 by category	
Asymptotic Distribution	$X^2\ follows\ chi - square\ (\chi^2)$ The sampling distribution is the same as the chi-square distribution as the expected frequencies become larger(infinite)	
Power	Use when no alternative exists and expected frequencies per cell are greater or equal to 5. Usually insensitive to ordering	

E. Multinomial Proportions Confidence Intervals

The ***Multinomial Proportion Confidence Intervals*** apply to cases where there are 2 or more possible responses. Let n_j denote the observed cell frequencies on j^{th} cell of m-cell multinomial distribution with the sample size n. The Multinomial distribution is the parameters $\pi_1, \pi_2, \ldots, \pi_m$. Each π_i denotes the probability that an observation will falls in the ith cell.

The Multinomial distribution is described in the probability mass functions module of *AroniSmartLytics*.

Each π_i can be estimated by its maximum likelihood estimator, $\hat{p}_i = \frac{n_j}{N}$, for j =1, 2, …,m. Hence π=($\pi_1, \pi_2, \ldots, \pi_m$)' can be estimated by $\hat{p} = (\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_m)'$. As the sample size becomes larger, $\sqrt{N(\hat{p}-\pi)}$ has m-variate normal distribution with zero mean vector and variance matrix ∑ with elements

$$\sigma_{jj} = \pi_i(1-\pi_i) \; for \; j = 1,2,\ldots,m$$

$$\sigma_{ij} = -\pi_j\pi_k \; for \; j \neq k$$

There are alternative multiple solutions proposed to estimate multinomial intervals, among the most commonly used being:

- Gold, Quesenberry, and Hurst Method
- Baily Method

The **Gold, Quesenberry and Hurst** approach uses two asymptotic simultaneous confidence intervals for πj:

$\hat{p}_j \pm \chi\left[\frac{\hat{p}_j(1-\hat{p}_j)}{N}\right]^{1/2}$ and $\frac{\chi^2+2n_j \pm \{\chi^2[\chi^2+2n_j(N-2n_j)/N]\}^{1/2}}{2(N+\chi^2)}$

Where χ^2 is the chi-square distribution with 1 degree of freedom estimated at 100(1- α/m) percentile, based on the Bonferroni inequality.

Baily proposed the following two simultaneous confidence intervals for πj:

$\left\{sin\left[Y_{1j} \pm \frac{\chi}{\sqrt{4N+2}}\right]\right\}^2$ and $\frac{\left\{Y_{2j} \pm \left[C(C+1-Y^2{}_{2j}\right]^{1/2}\right\}^2}{(C+1)^2}$

$for \; j = 1,2,\ldots,m$ where $C = \chi^2/4N$

and $Y_{1j} = sin^{-1}\sqrt{n_j + 3/8)/(N+3/4)}$

and $Y_{2j} = \sqrt{n_j + 3/8)/(N+1/8)}$ and χ^2 is the chi-square distribution with 1 degree of freedom estimated at 100(1- α/m) percentile, based on the Bonferroni inequality.

F. Poisson Confidence intervals and Test

The estimation of ***Poisson Confidence Intervals*** for the population parameter, the Poisson rate λ, $1/\lambda$ being the time interval between two Poisson events, with was proposed in V. Guerriero, A. Iannace, S. Mazzoli, M. Parente, S. Vitale, M. Giorgioni (2009). "Quantifying uncertainties in multi-scale studies of fractured reservoir analogues: Implemented statistical analysis of scan line data from carbonate rocks". Journal of Structural Geology (Elsevier).

The method uses the relationship between the Poisson and Normal distribution as shown in ***AroniSmartLytics™*** module on the relationships among the main probability distributions.

The method assumes a sample size between 15 and 20 observations.

Let n be the number of sampled points or events or observations and L the time interval.

The Poisson upper and lower limits of the 1-α confidence interval are given by:

$\lambda_{low} = \frac{\left(1-\frac{Z_{1-\alpha/2}}{\sqrt{n-1}}\right)n}{L}$ and $\lambda_{upp} = \frac{\left(1+\frac{Z_{1-\alpha/2}}{\sqrt{n-1}}\right)n}{L}$

where $Z_{1-\alpha/2}$ is the $1-\frac{\alpha}{2}$ percentile of the standardized normal distribution

G. Test of Homogeneity of Poisson Rates

A ***Poisson process may be homogenous or non-homogeneous.*** In homogenous Poisson processes, the rate parameter λ is assumed constant and does not vary with time or space. A non-homogenous Poisson process is a Poisson process in which the rate parameters, $\lambda(t)$ are not constant, but are a function of time and hence may vary with time. Researchers and practitioners are hence usually required to test whether the rates are homogenous before making inference.

The standard method of testing the homogeneity of a set of k Poisson frequencies is to apply the Poisson index of dispersion. The method usually uses the chi-square goodness of fit. With the method, the Poisson frequencies constitute observed frequencies in the k cells with expected values, with k >=2.

Consequently, Test of Homogeneity of Poisson Rates is a special case of Chi-Square goodness of fit.

H. Kolmogorov-Smirnov One-Sample Test

The ***Kolmogorov-Smirnov One-Sample*** test, named after the Soviet Russian mathematicians Andrey Nikolaevich Kolmogorov and Vladimir Ivanovich Smirnov, is suited when one is interested in the degree of agreement between the distribution of a set of observed sample values and some specified theoretical distribution. The researcher, student or practitioner is usually interested in testing whether the observed values could be reasonably assumed to have come from a population described by the theoretical distribution. The Kolmogorov-Smirnov test is also a kind of goodness-of-fit test.

Let F_0 be a completely specified cumulative relative frequency distribution function for the theoretical distribution under H_0.

Statistic/ Concept	Formula	Comment
Test / Variable Definition	$D_n = \max\left(\lvert F_0(X_i) - S_n(X_i)\rvert\right)$ $i = 1,2 \ldots n$ n= number of sample observations $F_0(X_i) = \frac{1}{n}\sum_{i=1}^{n} I_{X_i \le x}$	$F_0(X_i)$=proportion of cases expected to have a score of equal to or less than X_i $S_n(X_i)$=proportion of observed cases with a score of equal to or less than X_i $I_{X_i \le x}$ is an indicator function equal to 1 if $X_i \le x$ and 0 otherwise
Common Hypothesis	$Ho{:}\, D_n = 0.$ Kolmogorov Smirnov test focuses on the largest of the deviations between $F_0(X_i)$ and $S_n(X_i)$ Ho assumes D_n is too small and within the range of a random error	
Small samples	Usually expected frequency is at least 5 by category	
Asymptotic Distribution	$\sqrt{n}D_n$ follows the Kolmogorov ditribution $K = Max_{t\epsilon[0,1]}\lvert B(F(t))\rvert$, where B(F(t)) is the Brownian bridge. $\Pr(K \le x) = \frac{\sqrt{2\pi}}{x}\sum_{i=1}^{\infty} e^{-(2i-1)^2\pi^2/(8x^2)}$	
Power	More powerful than chi-square for small samples Assumes ordinal scale measurement	

I. Shapiro–Wilk test

The ***Shapiro-Wilk test***, named after the Canadian statisticians Samuel Shapiro and Martin Wilk , is used to asses the goodness of fit for a normal distribution. The Statistic assumes that under the null hypothesis, observed values come from a normally distributed population.

Statistic/ Concept	Formula	Comment
Test/ Concept Definition	$W = \frac{\left(\sum_{i=1}^{n} \alpha_i y_{(i)}\right)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$ *where* $y_{(i)}$ *is the ith smallest number in the sample or* ith *order statistic* $\bar{y}$ *is the sample mean* $= \sum y_i \ /n$ $(\alpha_{i,..,}\alpha_n) = \frac{m^T V^{-1}}{(m^T V^{-1} V^{-1} m)^{1/2}}$ where $(m_{i,..,}m_n)^T$ with $m_{i,..,}m_n$ the expected values of the order statistics of independent and identically-distributed random variables sampled from the standard normal distribution, and V is the covariance matrix of those order statistics.	
Hypothesis Testing and Test Validity	*The null hypothesis is rejected when W is too small*	
Other Assumptions and	The goodness of fit of a normal distribution may also be checked graphically via the quantile, Q-Q, plot	

J. Hodges-Lehmann Confidence Interval for The Median

The ***Hodges-Lehmann Confidence Interval for The Median test***, also known Hodges-Lehmann estimator or Hodges-Lehmann-Sen estimator, named after the American statisticians Joseph Hodges and Erich Lehmann and Indian statistician Kumar Sen, is used to estimate the location parameter or median of a population or test the differences between two population location parameters.

K. Lilliefors test

The ***Lilliefors test***, named after the American statistician Hubert Lilliefors is used to test the goodness-of fit under the null hypothesis of a normally distributed population. It is an extension of Kolmogorov–Smirnov test.

Lilliefors Test is similar to the Kolmogorov-Smirnov Test for one sample. However, unlike the Kolmogorov Smirnov test, Lilliefors test does not make any assumptions about the shape and the form of the normal distribution. The expected value or mean and the variance are not specified a priori, but estimated from the data. Hence, Lilliefors Test may be viewed as biased by the sampling error since the expected value and variance are estimated from the observed data.

Lilliefors involves the following steps:

- Step 1: Estimate the population mean and population variance from the observations.
- Step 2: Find the maximum discrepancy between the empirical (observed) distribution function and the cumulative distribution function (CDF) of the normal distribution with the estimated mean and estimated variance.
- Step 3: Just as in the Kolmogorov–Smirnov test, this will be the test statistic.

Researchers and practitioners often replace Lilliefors Test by the Q-Q plot.

L. Test for Distributional Symmetry

The *Test for Distributional Symmetry* is a test of central tendency. It is suited when one is interested in the shape of a distribution. The statistician wants to make inference on an unknown but symmetrical distribution. The test examines three subsets of three variables each: *Right triple*, *Left Triple* and *Neither*. Each of the possible triples is then coded as **left**, **right** and **neither** before computing the statistic of interest.

Statistic/ Concept	Formula	Comment
Test/Variable Definition	$T = \#right\ triples - \#left\ triples$	
Expected Value	Expected value of the standard normal distribution	
Variance	Variance of the normal distribution	
Common Hypothesis	$Ho{:}\mu_T = 0$ or X has a symmetric distribution. Under Ho, T is too small and within the range of a random error.	
Small samples	Usually expect sample size is around 20	
Asymptotic Distribution	z= T/σ_T $\sigma^2{}_T = \frac{(n-3)(n-4)}{(n-1)(n-2)}\sum_{i=1}^{n} {B_i}^2 + \frac{N-3}{N-4}\sum_{1\le j\le k\le n} B_{jk}^2 + \frac{n(n-1)(n-2)}{6} - \left[1 - \frac{(n-3)(n-4)(n-5)}{n(n-1)(n-2)}\right]T^2$	Z follows a standard normal distribution B_i= # right triples involving X_i -#left triples involving X_i B_{jk}= # right triples involving both X_j and X_k - # left triples involving both X_j and X_k
Power	Reasonably good for relatively large samples	Assumes interval scale measurement

M. The One-Sample Runs Test of Randomness

The ***One-Sample Runs Test of Randomness*** is a test of independency. It attempts to test whether a sample is randomnly drawn from a population, that is whether successive observations are independent.

Techniques exist that use the order of sequence of observations to test randomness. The One-Samples Runs Test of Randomness is based on a different approach which relies on runs. **A run** is a succession of identical symbols followed by different or no symbols.

For example a series of successive YES answer on a given question, before a NO answer is observed constitutes a run. The test may be one or two-tailed.

Statistic/ Concept	Formula	Comment
Test/Variable Definition	$r = \text{number of runs in the sample}$ n = sample size = m+k	m= # of observations in the one category k= # of observations in the other category
Common Hypothesis	$Ho: r_L < \mu_r < r_H$ or the observations appear in a random order. If the number of runs is less or equal to the lower threshold or greater or equal to the upper threshold, Ho is rejected	
Small samples	The test is suited for small samples (m and k less than 20)	
Asymptotic Distribution	Mean=$\mu_r = \frac{2mk}{n} + 1$ $\sigma^2{}_T = \frac{2\text{mk}(2\text{mk} - \text{n})}{(n-1)n^2}$ $z = \frac{r-\mu_r}{\sigma_r}$ $= \frac{r + h - \frac{2mk}{n} - 1}{\sqrt{[2mk(2mk-n]/[n^2(n-1)]}}$	$h = \begin{cases} \frac{1}{2} \text{ if } r < 2mk/n + 1 \\ -\frac{1}{2} \text{ if } r > \frac{2mk}{n} + 1 \end{cases}$ Z follows a standard normal distribution
Power-efficiency	Nonparametric test of randomness of a sequence of events.	

N. The Change-Point Test

The ***Change Point Test*** is a quantile test. It attempts to test whether there was a shift in an ordered sequence of observations. The test assumes that there is an underlying process that generates the observations in an ordered sequence, that the distribution of responses has one median, and at some point there was a shift in the median.

The test may be one or two-tailed. One-tailed test looks at upward or downward shift in the distribution. **Two-tailed** test assesses the potential shift, without regard to the direction of change.

Statistic / Concept	Formula	Comment
Test/ Variable Definition	For a binomial variable, X The number of successes: $m = \sum_{i=1}^{n} X_i$ The number of failures: k=n-m Cumulative number of successes at each point j in a sequence: $S_j = \sum_{i=1}^{j} X_i$ The largest absolute difference observed in the sequence: $D_{m,k} = max\left\|\frac{n}{mk}\left(S_j - \frac{jm}{n}\right)\right\|$	For an ordinal variable, X The sum of the ranks of the variables at or before point j: $W_j = \sum_{i=1}^{j} r_i$; j< n The statistic dividing the sequence into observations occurring before, m, and those occurring after k, the change: $K_{m,k} = max\left\|2W_j - j(n+1)\right\|$
Asymptotic Distribution	Variance of W_j: $\sigma^2{}_W = \frac{\mathrm{mk(n+1)}}{12}$ $z = \frac{W+h-\frac{m(n+1)}{2}}{\sqrt{[mk(n+1)]/12}}$ W is the sum of W_j where $K_{m,k}$ is maximized.	$h = \begin{cases} \frac{1}{2}\ if\ W < m(n+1)/2 \\ -\frac{1}{2}\ if W > \frac{m(n+1)}{2} \end{cases}$ Z follows a standard normal distribution
Power-efficiency	There is usually not direct and straightforward nonparametric for dichotomous variable The test is appropriate for ordinal or continuous measurement	

O. The McNemar Change Test

The ***McNemar Change Test***, named after the American psychologist and statistician Quinn McNemar, attempts to test the significance of changes after an experiment or event for subjects in a sample. In McNemar test, each subject serves as its own control.

The test assumes that there is measurement, at nominal or ordinal scale, on the subject "before" and "after" the experiment starts. A fourfold table crossing **After** and **Before** the event measurements is used in the study. The total number of subjects with changes is computed and used to test the significance of changes, without regard to the direction of change. Assuming that A subjects have their responses changed in one direction and B subjects have their responses changed in the other direction:

Statistic/ Concept	Formula	Comment
Test/ Variable Definition	$X^2 = \sum_{i=1}^{2} \frac{(O_i - E_i)^2}{E_i}$ $= \frac{[A-(A+D)/2]^2}{(A+D)/2} + \frac{[D-(A+D)/2]^2}{(A+D)/2}$ $= \frac{[A-D]^2}{A+D}$; with df=1	O_i= observed number of cases for category i E_i= expected number of cases for category I when Ho is true A+D total number of subjects whose response changed
Common Hypothesis	Ho: X^2 *follows chi − square* (χ^2) *distribution with degrees of freedom* $df = 1$. When Ho assumes equal proportion of cases in each category: $E_i = n/k$	
Small Expected Frequencies	Use the binomial test for very small samples	
Asymptotic Distribution	X^2 *follows chi − square* (χ^2) with df=1. The sampling distribution is the same as the chi-square distribution as the expected frequencies become larger(infinite	
Correction forContinuity	$X^2 = \frac{[\lvert A-D \rvert -1]^2}{A+D}$ with df=1	
Power	Use when samples are relatively larger. Otherwise use the binomial test	

P. The Sign Test

The *Sign Test* attempts to test the direction of the changes for two measures, especially when the measures are qualitative than quantitative. The test is suited for two related samples when the experimenter and researcher are only interested in the direction of the changes, with the assumptions of a progressive scale measurement.

There are no assumptions on the form and the shape of the underlying distribution or the populations from which the sample are drawn. The only requirement is the matching of the subjects under study. Assuming that A subjects have their responses changed in one direction and B subjects have their responses changed in the other direction:

Statistic/ Concept	Formula	Comment
Test/ Variable Definition	$P[X_i > Y_i] = P[X_i < Y_i] = 1/2$	X_i= judgment or score under one condition Y_i= judgment or score under the other condition $X_{i\,a} and\, Y_i$ are the matched scores
Common Hypothesis	Ho: $P[X_i > Y_i] = P[X_i < Y_i] = 1/2$	Ho assumes equal proportion of cases in each category
Small Samples	Use binomial test for very small samples with p=q=1/2	
Asymptotic Distribution/ Large Samples	Mean: $\mu_x = np = \frac{n}{2}$ Variance: $\sigma_x^2 = npq = n/4$ $z = \frac{2x - n}{\sqrt{n}}$	The sampling distribution Z is a normal distribution with mean n/2 and variance n/4
Correction for Continuity	$z = \frac{2x \pm 1 - n}{\sqrt{n}}$	This corresponds to using x+0.5 when x<n/2 and x-0.5 when x>n/2
Power	Use when samples are very small.	

Q. Marginal Homogeneity test

The ***Marginal Homogeneity Test*** attempts to test the significance in the differences among the rows and columns of a contingency table. It is often viewed as an extension of McNemar test.

Marginal homogeneity refers to equality (lack of significant difference) between one or more of the row marginal proportions and the corresponding column proportion(s).

When the rows and columns proportions are equal, there is agreement among the column and row factors. Marginal homogeneity is often useful in analyzing rater agreement.

Usually the contingency table is as follows, with Rater 1 and Rater 2 ratings subjects with scores 1, 2, ...k. Pij is the proportion or the frequencies of all observations assigned to score i on Rater 1 and j from Rater 2:

	Rater 2				
Rater 1	1	2	...	m	Column
1	p11	p12		p13	p1.
2	p21	p22		p23	p2.
...	...	...		...	...
m	p31	p32		p33	pr.
Row	p.1	p.2		p.k	1.0

Here p_{ij} denotes the proportion of all cases assigned to category i Rater 1 and category j by Rater 2.

The terms p1., p2., and pm. denote the marginal proportions for Rater 1--i.e. the total proportion or frequency of times Rater 1 uses categories 1, 2, .., m, respectively.

Similarly, $p_{.1}$, $p_{.2}$, and $p_{.m}$ are the marginal proportions for Rater 2.

The **Marginal Proportions** test can then be conducted using the Chi-square test, in case of assumed independence or McNemar test, if the observations are assumed to be potentially correlated.

R. Likelihood Ratio Test

The **Likelihood Ration Test** is used to compare the adequacy and fit of two models, either related or not. For related models, one is nested within and is a special case of the other. **The nested simpler model is the null hypothesis model**.

The test is based on the likelihood function. The likelihood function expresses the probability of some observed outcomes given specified parameter values. The likelihood function quantifies the likelihood of observations coming from a specified model. The likelihood ratio then expresses how many times more likely the data are under one model than the other.

For practical purposes and simplicity, the logarithm of the likelihood ratio is used. Consequently, the statistic is known as "**Log-likelihood ratio"**. The estimation of the statistic and the probabilities associated with the statistic can be done using ***Wilks' theorem or Neyman-Pearson Lemma***.

Based on *Wilk's therorem*, named after the American statistician Samuel Stanley Wilks, the statistic from the Log-likelihood ratio, usually denoted *D*, as the difference between the *2ln (likelihood)* approximately follows a chi-square distribution with degrees of freedom equal *df2-df1*, under the null hypothesis that the nested model is correct.

df2 and *df1* are respectively the number of free parameters from the **alternative (general)** and the **null (nested) model.**

Statistic/ Concept	Formula	Comment
Test/ Concept Definition	$D = -2ln\left(\frac{likelihood\ for\ the\ null\ (nested) model}{Likelihood\ foe\ the\ Althernative (general) model}\right)$ =-2ln(likelihood for the null model)+ 2ln(likelihood for the alternative model) $= 2[\ln(L_2) - \ln(L_1)]$	
Hypothesis Testing	$D = 2\ln(L_2/L_1) = 2[\ln(L_2) - \ln(L_1)]$ follows $\chi^2{}_{df2-df1}$	
Other Assumptions and	The Likelihood Function needs to be estimated in order to apply the test.	

S. Likelihood Ratio Confidence Intervals

The ***confidence intervals of the Likelihood Ratio*** Test are based on ***Neyman-Pearson Lemma***, named after the Polish American statistician Jerzy Neyman and British statistician Ego Sharpe Pearson.

From the Lemma:

Ho: null model with parameters θ_0 vs. H1: Alternative model with parameters θ_1

The likelihood ratio test which rejects Ho in favor of H1 is formulated as the most powerful test of size α when $\Lambda\,(\mathrm{x}) = \frac{L(\theta_0|x)}{L(\theta_1|x)} \leq \eta$ where $P(\Lambda\,(\mathrm{x}) \leq \eta|Ho) = \alpha$, the significance level.

Statistic/ Concept	Formula	Comment
Test/ Concept Definition	$\Lambda\,(\mathrm{x}) = \frac{L(\theta_0|x)}{L(\theta_1|x} \leq \eta$, where $L(\theta_0|x)$ is the likelihood function	
Hypothesis Testing and Test Validity	$q.P(\Lambda = c|Ho) + P(\Lambda < c|H0) = \alpha$ If Λ > c, do not reject H0; If Λ < c, reject H0; Reject with probability q if Λ = c.	
Other Assumptions and	The Likelihood Function needs to be estimated in order to apply the test.	

T. Confidence Intervals for Odds ratio

The **Confidence Intervals for Odds Ratio** are applied to assess the odds of an event or observation occurring in one group against the odds of the event or observation occurring in another group.

The odds ratio is used to measure the association or the degree of independence between two categorical or dichotomous samples or groups. In general, the problem suitable to Odds ratio tests is presented as follows:

,

	Variable Y		
Variable X	0	1	Column
0	p_{00}	p_{01}	$p_{1.}$
1	p_{10}	p_{11}	$p_{2.}$
Row	$p_{.1}$	p.2	1.0

where p11, p10, p01 and p00 are non-negative "cell probabilities" that sum to one. The contingency table from the observations is:

	Y = 0	Y = 1
X = 0	n_{00}	n_{01}
X = 1	n_{10}	n_{11}

With the estimated proportions, $\hat{p}_{ij}$

where $\hat{p}_{ij}$= nij / n, with n = n11 + n10 + n01 + n00 being the sum of all four cell counts.

The odds for Y within the two subpopulations defined by X = 0 and X = 1 are defined in terms of the conditional probabilities given X:

	Y = 0	Y = 1
X = 0	p_{00} / (p_{00} + p_{01})	p_{01} / (p_{01} + p_{00})
X = 1	p_{10} / (p_{11} + p_{10})	p_{11} / (p_{10} + p_{11})

Thus the odds ratio is:

$$\text{odd ratio} = \frac{p_{11}/(p_{11}+p_{10})}{p_{10}/(p_{11}+p_{10})} \Big/ \frac{p_{01}/(p_{01}+p_{00})}{p_{00}/(p_{01}+p_{00})} = \frac{p_{11}p_{00}}{p_{10}p_{01}}$$

Statistic/ Concept	Formula	Comment
Test/ Concept Definition	The sample log odds ratio is: $\log\left(\frac{\hat{p}_{11}\hat{p}_{00}}{\hat{p}_{10}\hat{p}_{01}}\right) = \log\left(\frac{n_{11}n_{00}}{n_{10}n_{01}}\right)$	
Hypothesis Testing and Test Validity	The distribution of the log odds ratio approximately follows a normal distribution with mean = sample log odd ratio and standard deviation = $\sqrt{\frac{1}{n_{00}}+\frac{1}{n_{10}}+\frac{1}{n_{01}}+\frac{1}{n_{11}}}$	
Other Assumptions and	The odd ratios have a symmetrical property.	

U. Barnard's test

The *Barnard's test*, named after the British statistician George Alfred Barnard is used to test the independence of rows and columns in a contingency table. It is an exact test and presents an alternative to and with the claim of being sometimes more powerful than the **Fisher's Exact test**. Barnard's test is more computationally involving than the Fisher's exact test.

V. Savage and Maximum Efficiency Robust Test (MERT) Scores.

The Savage Test, named after the **American statistician Leonard Jimmie Savage**, uses ***the rank scores***, or ***linear rank statistic*** to test the independence of observations with the assumptions of a logistic distribution under the null hypothesis.

The Savage exact scores from a set of observations drawn from a sample of size n and ranked from lowest to the largest value are:

$\alpha^*[n,i] = \frac{1}{n} + \frac{1}{N-1} + \cdots + \frac{1}{n-i+1}; i = 1,..,n$

Asymptotically the Savage test corresponds to log-rank test.

$$b[n,i] = \ln \{1 - \frac{i}{n+1}\}; i = 1,..,n$$

Savage Scores test is popular in the area of Game theory economics, in which Leonard Jimmie Savage was a renown expert.

MERT: Maximum Efficiency Robust test, developed by **American statistician Joseph L. Gaswirth** is also related to log-rank test. It is a linear combination of efficient robust scores and is highly dependent on the situation under analysis. There are several theoretical derivations of MERT tests.

W. The Wilcoxon Signed Ranks Test

Experimenters are often interested not only in the direction, but also in the magnitude of changes. While the **Sig Test** is suitable for the direction of the changes, it does not test the magnitude.

Wilcoxon Signed Ranks, named after the **American chemist and statistician Frank Wilcoxon,** tests whether the means of the population, from which two paired samples are drawn, differ. The test is used to assess the magnitude by assigning weights to the paired observation with a large difference.

To identify the pairs with the large absolute difference, the observed differences are ranked from the smallest to the largest, without regard to sign (absolute value of the differences) with the smallest assigned rank 1. Two statistics are developed from the ranks:

T^+ = the sum of the ranks of the positive differences and

T^- = the sum of the ranks of the negative differences.

Statistic/ Concept	Formula	Comment
Test/ Variable Definition	$d_i = X_i - Y_i$ number of ranks = n(n+1)/2	X_i= observed value of pair i under treatment one Y_i= observed value of pair i under treatment two
Common Hypothesis	Ho: *Sum of the ranks of positive* d_i *is equal to the sum of the ranks of negative* d_i. Ho assumes that the two treatments are equivalent	
Small Samples	Use binomial test for very small samples with p=q=1/2	
Asymptotic Distribution/ Large Samples	Mean: $\mu_{T^+} = \frac{n(n+1)}{4}$ Variance: $\sigma^2_{T^+} = \frac{n(n+1)(2n+1)}{24}$ $z = \frac{T^+ - \mu_{T^+}}{\sigma^2_{T^+}} = \frac{T^+ - n(n+1)/4}{\sqrt{n(n+1)(2n+1)/24}}$	Z is approximately normally distributed with mean 0 and variance 1
Correction for Tied Ranks	$\sigma^2_{T^+} = \frac{n(n+1)(2n+1)}{24} - 1/4.8 \sum_{j=1}^{k} t_j(t_j - 1)(t_j + 1)$; where k= number of groupings of different tied ranks and tj = number of tied ranks in grouping j.	
Power	The test is powerful for small samples	

X. The Permutation Test For Paired Replicates

Experimenters are often interested in the occurrence of specific data within the observations. The **Permutation Test** is a powerful nonparametric test for this type of situations. It doesn't make any assumptions about a specific form of the distribution, normality, or homogeneity of the variance. The test is suited for all kinds of data that have a numerical meaning.

The test looks at the all possible pairing of the observations from two treatments, and assumes that each pairing is randomly assigned. The test may be one or two-tailed.

Statistic/ Concept	Formula	Comment
Test/ Variable Definition	$d_i = X_i - Y_i$ Number of possible outcomes for random pairs: 2^N Number of possible outcomes in the region of rejection: $\alpha 2^N$	X_i= observed value of pair i under treatment one Y_i= observed value of pair i under treatment two $\alpha = signicance\ level$
Common Hypothesis	$Ho\ for\ two-tailed\ test$: $Sum\ of\ differences\ of$ $positive\ d_i\ is\ equal\ to\ the$ $sum\ of\ the\ differences\ of\ negat$	Ho assumes that the two treatments are equivalent
Power	The test is powerful among non parametric tests, especially for ordinal and interval data	

Y. The Fisher Exact Test for 2x2 Tables

In experiments, often paired samples may be impractical, too costly, or inappropriate. In such cases, independents samples may be used instead.

The Fisher Exact Test for 2x2 Tables, named after the English statistician and biologist Sir **Ronald Aylmer Fisher**, is a very useful test when analyzing nominal or ordinal data with small independent samples. It is appropriate when observations from two independent samples fall into one of two mutually exclusive and specific classes. Every subject in each sample may fall in one or the other class based on the observation value.

Observations may be classified as in the following contingency table:

Variable	Group or Sample		Combined Frequency
	I	II	
Class 1	A	B	A+B
Class 2	C	D	C+D
Total	A+C	B+D	n

The test may be one or two-tailed.

Statistic/ Concept	Formula	Comment
Test/ Variable Definition	$p = \frac{\binom{A+C}{A}\binom{B+D}{B}}{\binom{n}{A+B}} = \frac{[(A+B)!\,(C+D)!\,(A+C)!\,(B+D)!]}{n!\,A!\,B!\,C!\,D!}$.	A, B, C, D= number of subjects in one category of sample vs treatment p is the exact probability of observing a specific set of frequencies in a 2x2 contingency table.
Common Hypothesis	$Ho:\ p_1 = p_2$ p_i= probability that an observation on randomnly selected subject from group i falls in a given class. Ho assumes that the assignment to the two groups is random. For two tailed tests, the alternative is that the assignment is not random. For one-tailed test, the alternative is whether there is a significantly higher proportion of one class.	
Asymptotic Distribution/ Large Samples	The test uses exact probabilities and makes no assumptions for asymptotic distribution	
Power	The test is powerful among non parametric tests, especially for dichotomous nominal	

Z. The Chi-square Test for 2 independent samples

For two independents samples or groups, when the data consists of frequencies in discrete categories, the **Chi-square Test for Two Independent Samples** may be the right test to use.

The experimenter wants to test whether the two groups differ with respect to some characteristics. If the groups do not differ with respect to the characteristics, then there is interaction between the groups and the variable of interest. The test compares the proportions of cases from one sample or group in the various categories of the variable. The decision about the independence between the groups and the variable of interest depends on the result of the comparison: if the proportions are the same across the groups for all the classes, then there is independence; otherwise there is interaction.

The test is accomplished based on a threshold of proportionality expected to occur as a chance when there is no interaction.

Statistic/ Concept	Formula	Comment
Test/ Variable Definition	$X^2 = \sum_{i=1}^{k}\sum_{j=1}^{c}\frac{(x_{ij}-E_{ij})^2}{E_{ij}} = \sum_{i=1}^{k}\sum_{j=1}^{c}\frac{{x_{ij}}^2}{E_{ij}} - n$ n = sample size k=number of groups c=number of categories of classes for the variable	x_{ij}= observed number of cases for the category or class i under treatment j E_{ij}= number of cases expected for the category or class i under the treatment j
Common Hypothesis	*Ho : There is independence between the two groups and the variale of interest.* Ho assumes that the two groups/samples/ treatments are independent of the classes of the variable or that there is no interaction between the groups and the variable of interest	
Small expected values	The chi-square test assumes a minimum of frequencies within each cell.	Usually a minimum of 5 subjects in each cell is expected.
Asymptotic Distribution/ Large Samples	X^2 *follows chi − square* (χ^2) with degree s of freedom df= (r-1)(c-1)	
2x2 Contingency Table and Correction for Continuity	$X^2 = \frac{n[\lvert AD-BC\rvert - n/2]^2}{(A+B)(C+D)(A+C)(B+D)}$ with df=1	
rx2 tables- Equation for partition t.	${X^2}_t = \frac{n^2\left(n_{t+1,2}\sum_{i=1}^{t} n_{i1} - n_{t+1,1}\sum_{i=1}^{t} n_{i2}\right)^2}{C_1 C_2 R_{t+1}\left(\sum_{i=1}^{t} R_i\right)\left(\sum_{i=1}^{t+1} R_i\right)}$ t =1,2..,r-1, C_i= sum for column i and R_i sum for row i	
Power	The test is usually the best where it is appropriate. It may not be appropriate for situations where order is taken into account	

AA. The Median Test

In experiments with two independents samples, experimenters may be interested in knowing whether the two groups *differ in central tendencies*. The **Median Test** helps to assess whether two independent groups, with or without the same size, have likely been drawn from populations with the same median.

The Median test is appropriate for nominal or ordinal scale measurements. There may also be specific instances where even interval measures may not have alternative to the Median Test. Observations may be classified as in the following contingency table:

Variable		Group or Sample I	Group or Sample II	Combined Frequency
Class 1		A	B	A+B
Class 2		C	D	C+D
	Total	m	r	n

Statistic/ Concept	Formula	Comment
Test/ Variable Definition	$p[A,B]=\frac{\binom{m}{A}\binom{r}{B}}{\binom{n}{A+B}}$ $p=\frac{[(A+B)!\,(C+D)!\,m!\,r!]}{n!\,A!\,B!\,C!\,D!}$	A, B, C, D= number of subjects in one category of pair i under treatment one; p is hypergeometric probability distribution.
Common Hypothesis	$Ho:\ median\ of\ Group\ I = median\ Group\ II.$ Ho assumes that the medians in the populations from which the two samples were drawn are the same. The test may be one or two-tailed.	
Asymptotic Distribution/ Large Samples	When n is large, use the chi-square (corrected for continuity). Otherwise use the Fisher Exact test	
Correction for Continuity	$X^2=\frac{n[\lvert AD-BC\rvert-n/2]^2}{(A+B)(C+D)(A+C)(B+D)}$ with df=1	
Power-Efficiency	The power efficiency decreases as the sample size increases reaching 2/π.	

BB. The Wilcoxon-Mann-Whitney Test

The ***Wilcoxon-Mann-Whitney***, also know as ***Mann-Whitney U*** test, is a nonparametric test that can be used to analyze data from a two independent groups design when the measurement is at least ordinal.

Named after **American statistician-chemist Frank Wilcoxon, Austrian-American statistician Henry Mann and American Statistician Donald Ransom Whitney**, it may help to analyze not only the independence, but also the degree of separation (or the amount of overlap) between the two groups. The test assumes that observations of the two groups are sampled from the same continuous distributions.

The null hypothesis assumes that the two sets of scores are samples from the same population, and hence, 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 one-tailed, it further specifies the direction of the differences, i.e., scores from the test group are systematically higher or lower than the scores from the control group.

To apply **Wilcoxon-Mann-Whitney** test, the scores from the two groups are ranked in order of increasing size, taking into account the algebraic size, i.e, assigning the lowest ranks to the largest negative values. The ranks from each group are then summed separately.

Statistic/ Concept	Formula	Comment
Test/ Variable Definition	$W_x + W_y = \frac{N(N+1)}{2}$ n=sample size for X. m= sample size for Y. N= sample size = m+n	W_x= sum of the ranks from group X W_y= sum of the ranks from group Y

Common Hypothesis	*Ho*: the average rank in the two groups are equal	If the average rank in one group is greater than the average in the other group, the Ho is rejected,
Small Samples	For small samples the statistic determines the exact probabilities under Ho	
Asymptotic Distribution/ Large Samples	Mean: $\mu_{W_x} = \frac{m(N+1)}{2}$ Variance: $\sigma^2{}_{W_x} = \frac{mn(N+1)}{2}$ $z = \frac{W_X \pm .5 - \mu_{W_X}}{\sigma_{W_X}} = \frac{W_X \pm .5 - m(N+1)/2}{\sqrt{mn(N+1)/12}}$	Z is asymptotically normally distributed with zero mean and unit variance. 0.5 is added or deducted for finding respectively left and right tail probability
Correction for ties	$z = \frac{W_X \pm .5 - m(N+1)/2}{\sqrt{[mn/N(N-1)]\left[(N^3-N)/12 - \sum_{j=1}^{g}\left(\sum_{j=1}^{g} t^3 - t_j\right)/12\right]}}$ with g the number of groupings of different tied ranks and t_j the number of tied ranks in the jth grouping.	
Power-efficiency	The test may be a powerful alternative to t test, with power-efficiency reaching up to 3/π as N increases.	

CC. Robust Rank-Order Test

The ***Wilcoxon-Mann-Whitney*** test assumes that observations of the two independent groups are sampled from the same continuous distributions and are on at least the ordinal scale. Although the null hypothesis states that the medians from the two groups are the same, it is implied that the distributions are the same, hence the variances of the distributions are equal. The alternative assumes that the variances are the same, but only the medians systematically differ.

In some instances, **the experimenter wants to test whether the medians are the same without assuming that the distributions are the same.** The Wilcoxon-Mann-Whitney may not be appropriate in these instances. The **Robust Rank-Order** Test is the best alternative for such cases, especially in situations known as the **Behrens-Fisher** problem, where measurements are restricted in range and based on other factors.

To **apply Robust Rank-Order test**, the scores from the two groups are ranked in order of increasing size, taking into account the algebraic size, that is, assigning the lowest ranks to the largest negative values.

For each value in the control group X, the number of observations of the test group Y with a lower rank is captured. This number, called "placement" of the X scores is denoted $U(YX_i)$.

Then the mean of $U(YX_i)$ is computed and denoted $U(YX)$. Similarly the placement of Y scores, denoted $U(XY_j)$, the number of observations from X that precede each Y_j is tabled, and the mean $U(YX)$ calculated.

Statistic/ Concept	Formula	Comment
Test/ Variable Definition	$\grave{U} = \frac{mU(YX) - nU(XY)}{2\sqrt{V_x + V_y + U(XY)U(YX)}}$ where: $U(YX) = \sum_{i=1}^{m} \frac{U(YX_i)}{m}$ $U(XY) = \sum_{j=1}^{n} \frac{U(XY_i)}{n}$ n=sample size for X m= sample size for Y	V_x= index of variability for group X V_y= index of variability for group Y $V_x = \sum_{i=1}^{m} [U(YXi) - U(YX)]^2$ $V_y = \sum_{j=1}^{n} [U(XYi) - U(XY)]^2$
Common Hypothesis	*Ho*: there is no difference between the two groups with regard to the variable of interest	If the observed value of $\grave{U}$ has an associated probability less or equal to the threshold value, reject Ho,
Small Samples	For small samples the statistic determines the probabilities under Ho	
Power-efficiency	Similar power-efficiency as the Wilcoxon test but suited for testing the independence without requiring equal variance.	

DD. Kolmogorov-Smirnov Two-Sample Test

The ***Kolmogorov-Smirnov Two-Sample*** test, named after the **Soviet Russian mathematicians Andrey Nikolaevich Kolmogorov** and **Vladimir Ivanovich Smirnov**, is suited when one is interested in knowing whether two independent samples were drawn from the same population or two populations with the same distribution.

The researcher, student, or practitioner is usually interested in testing whether there is an agreement between the cumulative distributions of the two samples.

Let S_m and S_n be the cumulative relative frequency distribution function for, respectively, the first sample with size ***m*** and second sample with size ***n***.

In the case of two samples, the differences may come from deviations dues to central tendencies, skewness, dispersion, etc.

Hence two tests are usually used: one tailed and two-tailed test. Two-tailed tests looks at the differences among the distributions, whereas one-tailed test looks at the direction of the difference.

Statistic/ Concept	**Formula**	**Comment**
Test/ Variable Definition	$D_{m,n} = \max(\lvert S_m(X) - S_n(X)\rvert)$ for two tailed test $D_{m,n} = \max([S_m(X) - S_n(X)])$ for one-tailed test n, m= number of sample observations	$S_m(X)$: proportion of observed cases from sample m with a score of equal to or less than X $S_n(X)$: proportion of observed cases from sample n with a score of equal to or less than X
Common Hypothesis	$Ho: D_{m.n} = 0.$ Alternative H1: **Two tailed:** the population are from different populations. **One-tailed:** the population values from which one sample was drawn are stochastically larger than the population values from which the other sample was drawn.	Reject Ho when $\sqrt{\frac{mn}{m+n}} D_{mn} > K_\alpha$

Small/Large samples	Usually use $mnD_{m.n}$ for small samples and $D_{m.n}$ for large samples	Assumes ordinal scale measurement
Asymptotic Distribution	$X^2 = 4D^2{}_{m,n}\frac{mn}{m+n}$ is approximated by the chi-square distribution with 2 degrees of freedom (df=2)	
Power-Efficiency	Higher power-efficiency than that of t-test for small samples. Power decreases with larger samples. Higher power-efficiency than chi-square and median test.	

EE. Anderson-Darling Test

The ***Anderson-Darling*** test, named after the named after **American mathematical statisticians Theodore Wilbur Anderson** and **Donald A. Darling**, is used to assess whether a given sample of data did not arise from a given probability distribution. The test is distribution free and assumes that there are no parameters to be estimated in the distribution being tested.

It is most often used in cases where experimenters are interested in a family of distributions. In those cases, the parameters of the family of distributions along with the critical values are estimated from the data and the test needs to be adjusted accordingly. The test can be viewed as a goodness of fit test and will be used to test normality and departures from normality. The test is also often used in parameter estimation, by estimating and minimizing distances.

Anderson-Darling Test is available for both one and k-samples. In k-sample settings, the goal is to test whether several collections of observations can be modeled as coming from a single population, where the distribution function does not have to be specified.

For **one sample-test**, the Anderson–Darling test assesses whether a sample comes from a specified distribution. It makes use of the fact that with a given hypothesized underlying distribution, under the null hypothesis that the data arise from this distribution, the data can be transformed to a uniform distribution.

The test is conducted against the cumulative distribution of the hypothesized distribution. Hence the data is first ordered in ascending order: *{$Y_1 < Y_2 < Y_3$ --- Y_n}*.

Statistic/ Concept	Formula	Comment
Test/ Variable Definition	$A^2 = -n - S$ where $S = \sum_{k=1}^{n} \frac{2k-1}{n}[lnF(Y_k) + \ln(1 - F(Y_{n+1-k}))]$	$F(\ )$ is the cumulative density or mass function of the hypothetical probability distribution. No parameter from F is estimated.
Normality	*Test for Normality*: $$Yi = \frac{X_i - \hat{\mu}}{\hat{\sigma}}$$ $$A^2 = -n - \frac{1}{n}\sum_{i=1}^{n}[(2i-1)ln\Phi(Y_i) + (2(n-i)+1)\ln(1-\Phi(Y_i))]$$ where $\hat{\mu} = \begin{cases} \mu, & \text{if the mean is known} \\ \bar{X} = \frac{1}{n}\sum_n^n X_i & \text{if the mean is unknown} \end{cases}$ $$\widehat{\sigma^2} = \begin{cases} \sigma^2, & \text{if the variance is known} \\ \frac{1}{n}\sum_n^n (X_i - \mu)^2 & \text{if known mean known and unknowo variance} \\ \frac{1}{n-1}\sum_n^n (X_i - \bar{X})^2 & \text{otherwise} \end{cases}$$	
Power-Efficiency	Powerful for testing normality.	Comparable to Kolmogorov-Smirnov and Shapiro-Wilk tests

FF. The Permutation Test For Two Independent Samples

The *Permutation Test* is a powerful nonparametric test for assessing the significance of the differences between the means of two independent samples, when the two sample sizes *m* and *r* are small. The test is suited for the data at the interval scale since it uses the values of the scores. The testing approach determines the exact probability associated with the observations without making assumptions about the underlying distributions in the populations of interest.

For testing, specify the region of rejection that includes the number of the most extreme possible outcomes. The extremes combinations are those that give the largest differences between the sum of values from the potential control sample $\sum X$ and the sum of values from the potential test sample, $\sum Y$.

Statistic/ Concept	Formula	Comment
Test/ Variable Definition	Number of possible outcomes: $\binom{m+r}{r}$ Number of possible outcomes in the region of rejection: $\alpha\binom{m+r}{r}$	m = number of observed values in the control treatment r= number of observed values on the test treatment
Common Hypothesis	*If the observed scores fall within rejection region, reject Ho at level α of significance*	Ho assumes that the two treatments are equivalent
Asymptotic Distribution/ Large Samples	$t = \frac{\hat{X}-\hat{Y}}{\sqrt{\sum(X_i-\hat{X})^2/(m(m-1)+\sum(Y_i-\hat{Y})^2/(mn(r(r-1)}}$, has approximately a Student t distribution with df= m+r-1	
Power-Efficiency	The test is powerful among non parametric tests	

GG. The Siegel-Tukey Test for Scale Differences

*The **Siegel–Tukey test***, named after the **American psychologist Sidney Siegel** and **American statistician John Tukey**, is a non-parametric test which may be applied to data measured at least at an ordinal scale to test for differences in scale and variability between two groups.

The test is used to determine if one of two groups of data tends to have more widely dispersed values than the other. In other words, the test determines whether one of the two groups tends to move, sometimes to the right, sometimes to the left, but away from the center.

The test is suited for the data at the ordinal scale, when one group is expected to have more variability that the other.

The ***Siegel-Tukey test*** focuses on the range or spread of one group compared to another. The test assumes that the medians of the two groups are the same or known. If they are known, they are subtracted from the observation values to adjust the scores, so that the medians are equal.

To compute the test, combine the observations from the *X* and *Y* groups and arrange them in a single ordered series, associating each observation with its group. **Assign ranks order to observations sequentially and alternatively from the most extreme observations to the lowest.** The approach assumes that the extreme observations are atypical and the central (lowest scores) are typical.

Finally calculate the sums of the ranks for each group. Testing is conducted using Wilcoxon test.

Statistic/ Concept	Formula	Comment
Test/ Variable Definition	W_x =the sum of the ranks in group X adjusted for sample size W_y =the sum of the ranks in group Y, adjusted for sample size.	m = number of observed values in the control treatment n= number of observed values on the test treatment
Common Hypothesis	Ho: *the dispersion of the two groups is the same* that is $H_0: \sigma_x^2 = \sigma_y^2$ $H_1: \sigma_x^2 > \sigma_y^2$	Ho assumes that the dispersion among the two groups are equivalent, hence Wx and Wy are about the same. The probability comes from Wilcoxon test with **m** and **n** for Wy value
Asymptotic Distribution/ Large Samples	N/A	
Power-Efficiency	The test usually has a relatively low power-efficiency	

HH. The Moses Rank-Like Test for Scale Differences

The ***Siegel–Tukey test*** is used to test for differences in scale and variability between two groups. Hence, the test helps assess homogeneity and heterogeneity between two groups. However, the test assumes that the medians are the same or known.

When the medians are assumed not equal and unknown, the Siegel-Tukey test is not applicable. The ***Moses Rank-Like Test for Scale Differences*** is used in these cases. The test also focuses on the range or spread of one group compared to another, when observations are measured at least at the interval scale.

To compute the test, the observations from the X and Y groups are divided into equal subsets of at least two observations in each group. The assignment of observations to a subset must be random. Observations not assigned to a subset are discarded. For each subset, calculate the sums of the differences from the subset mean. Then, for each group, compute the dispersion.

Statistic/ Concept	Formula	Comment
Test/ Variable Definition	$D(X_j) = \sum_{i=1}^{k}(X_{ij} - \bar{X}_j)^2$, dispersion index for group X $D(Y_j) = \sum_{i=1}^{k}(Y_{ij} - \bar{Y}_j)^2$, dispersion index for group Y where $\bar{X}_j = {\sum_{i=1}^{k} X_{ji}}/{k}$ and $\bar{Y}_j = {\sum_{i=1}^{k} Y_{ji}}/{k}$	$m =$ number of subsets of X n= number of subsets of Y X_{ji} data for subset j of X, Y_{ji} data for subset j of Y, i=1,2,..,k; j=1,..,m or n. k= number of data in a subset.
Common Hypothesis	Ho: *the dispersion of the two groups is the same* that is $H_0: \sigma_x^2 = \sigma_y^2$ $H_1: \sigma_x^2 \neq \sigma_y^2$ for two-tailed test Ho assumes that the variability among the two groups is the same. $H_1: \sigma_x^2 > \sigma_y^2$ or $H_1: \sigma_x^2 < \sigma_y^2$ for one-tailed test	
Asymptotic Distribution /Large Samples	For large samples, the approximation of the Wilcoxon test may be used	
Power-Efficiency	The efficiency increases with the size of the subsets and the sample size.	

II. The Cochran Q Test

The *Cochran Q* Test named after the **Scottish-American statistician William Gemmell Cochran** is used to analyze a two-way randomized block designs where the response variable can take only two possible outcomes (coded as 0 and 1, or Yes and No). It may be looked at as the extension of MacNemar test with multiple related samples, k in which k subjects, each from a different group are matched.

The Cochran test assumes that there are k > 2 experimental treatments and that the observations are arranged in b blocks as follows:

Groups	Sample or Treatment 1	2	…	k	Number of Successes
Block 1	X_{11}	X_{12}		X_{1k}	X_{1*}
Block 2	X_{21}	X_{22}		X_{2k}	X_{2*}
Block 3	X_{31}	X_{32}		X_{3k}	X_{3*}
…					
Block b	X_{b1}	X_{b2}		X_{bk}	X_{b*}
Total	X_{*1}	X_{*2}		X_{*k}	X_{**}

Statistic/ Concept	Formula Comment
Test/ Variable Definition	$Q = k(k-1)\frac{\sum_{i=1}^{k}\left(X_{*j}-\frac{X_{**}}{k}\right)^2}{\sum_{i=1}^{k}X_{i*}(k-X_{i*})}$ k= number of treatments or samples X_{*j}= the total for column (sample) j X_{i*}= the total for row (block) i b= the number of blocks or groups n= the grand total or sample size.
Common Hypothesis	Ho: *there is no differences in the scores across all groups*
Asymptotic Distribution/Large Samples	Q is approximated by χ^2 with df=k-1
Power-Efficiency	Cochran Q test is suited for dichotomous data where an alternative parametric statistic may not exist

JJ. Friedman Two-Way Analysis of Variance by Ranks

The Friedman Two-Way Analysis of Variance by Ranks, named after the **U.S. economist Milton Friedman** is a non-parametric statistical test to test whether k matched samples have been drawn from the same population. Similar to the parametric repeated measures Analysis of Variance (ANOVA), it is used to detect differences in treatments across multiple test attempts.

The procedure involves arranging the observations in a two-way table with n rows and k columns. The rows represent the matched sets of subjects, and the columns represent the groups, conditions or treatments.

The scores in each row (or block) are combined together and ranked from 1 to k. Then the values of ranks by columns are considered. The Friedman test determines the probability that the samples, that is, the different columns of ranks come from the same population, or equivalently that the k samples have the same median.

Applicable to complete block designs, it is thus a special case of the **Durbin test**.

Subjects	Sample or Treatment			
	1	2	..	k
1	X_{11}	X_{12}		X_{1k}
2	X_{21}	X_{22}		X_{2k}
3	X_{31}	X_{32}		X_{3k}
...				
n	X_{b1}	X_{b2}		X_{bk}
Sum of Ranks	X_{*1}	X_{*2}		X_{*k}

Statistic/ Concept	Formula	Comment
Test/ Variable Definition	$F_r = \left[\frac{12}{nk(k+1)}\sum_{j=1}^{k} r_{\cdot j}{}^2\right] - 3n(k+1)$ k= number of treatments or groups (columns) n= number of rows (subjects) r_{*j}= Sum of ranks in column j	b= the number of blocks, groups or treatments n= the sample size.
Common Hypothesis	$Ho: \theta_1 = \theta_2 = \cdots = \theta_k$ $H_1: \theta_i \neq \theta_j$, for at least two of the treatments	*There is no differences in the medians across all the treatments*
Asymptotic Distribution/ Large Samples	Fr is approximated by χ^2 with df=k-1	
Correction for Ties	$F_r = \frac{\left[12\sum_{j=1}^{k} r_{\cdot j}{}^2\right] - 3n^2k(k+1)^2}{nk(k+1) + \frac{nk - \sum_{i=1}^{n}\sum_{j=1}^{g_i} t_{ij}{}^3}{(k-1)}}$; where t_{ij} is the size of the jth set of tied ranks in the ith group and g_i= the number of sets of tied ranks in the ith group.	
Power-Efficiency	The power-efficiency of Friedman test may be compared to F test as k and n increase.	

KK. Quade Test

The Quade Test, also known as ***Quade ANOVA Test***, is used as an alternative of the Friedman Two-Way Analysis of Variance by Ranks. As the Friedman test is a k-sample extension of the sign test, the Quade test is an extension of the Wilcoxon signed-rank for paired or k related samples.

In Quade ANOVA test, the observations in each row of a two-way table are ranked. Then the range of the ranks for each row is found and the ranges of the rows are ranked. The Quade Test is computed based on the ranks of the row ranges.

Subjects	Sample or Treatment			
	1	2	..	k
1	X_{11}	X_{12}		X_{1k}
2	X_{21}	X_{22}		X_{2k}
3	X_{31}	X_{32}		X_{3k}
…				
n	X_{b1}	X_{b2}		X_{bk}
Sum of Ranks	X_{*1}	X_{*2}		X_{*k}

Statistic/ Concept	Formula	Comment
Test/ Variable Definition	$Q = \frac{(n-1)B}{A-B}$ where k= number of treatments or groups (columns) n= number of rows (subjects) $A = \sum\sum S_{ij}^2 \;\; B = \sum S_j^2 \;\; S_j = \sum S_{ij}$ $S_{ij} = Q_i\left[R_{ij} - \frac{k+1}{2}\right]$	$R_{ij,i=1,..,n,j=1,..,k}$ the ranks of the observations $Q_{i,i=1,..,n,j=1,..,k}$ the ranks of the row ranges
Common Hypothesis	$Ho: \theta_1 = \theta_2 = \cdots = \theta_k$ $H_1: \theta_i \neq \theta_j$, for at least two of the treatments	*There is no differences in the medians across all the treatments*
Asymptotic Distribution/ Large Samples	Qis approximated by F-distribution with df=k-1 and (n-1)(k-1) under the null hypothesis of independence	
Power-Efficiency	The power-efficiency of Quade Test is compared to Friedman Two Analysis of Variance by Ranks.	

LL. The Page Test for Ordered Alternatives

The Page Test for Ordered Alternatives is an extension of the Friedman test. It also known as **Page's Trend Test or Page's L test**. It tests whether the groups (or measures) are the same versus the alternative that the groups or measures are ordered in a specific sequence. The a priori order of the groups is required for the test. The Page test is useful where:

- **there are three or more conditions**, a number of subjects (or other randomly sampled entities) are all observed in each of them, and

- **there is a prediction** that the observations will have a **particular order.**

Statistic/ Concept	Formula	Comment
Test/ Variable Definition	$L = \sum_{j=1}^{k} jr._j$ k= number of treatments or groups (columns) r_{*j}= Sum of ranks in column j	b= the number of blocks, groups or treatments n= the sample size.
Common Hypothesis	$Ho: \theta_1 = \theta_2 = \cdots = \theta_k$ $H_1: \theta_1 \leq \theta_2 \leq \cdots \leq \theta_k,$	*There is no differences in the medians across all the treatments*
Asymptotic Distribution/ Large Samples	$\mu_L = \frac{nk(k+1)^2}{4}$ $\sigma^2{}_L = \frac{nk^2(k^2-1)^2}{144(k-1)}$ $z_L = \frac{12L - 3nk(k+1)^2}{k(k^2-1)} \sqrt{\frac{k-1}{n}}$	z_L is approximately normally distributed with zero mean and standard deviation one.
Power-Efficiency	Page test is more powerful than Friedman for the ability to detect ordered alternatives.	

MM. The Chi-square Test for k independent samples

The Chi-square Test for k independent samples is used to test the significance of differences among k samples for discrete categorical, nominal, or ordinal measurements. The test is an extension of the chi-square test for two independent samples.

Statistic/ Concept	Formula	Comment
Test/ Variable Definition	$X^2 = \sum_{i=1}^{r}\sum_{j=1}^{k}\frac{(x_{ij}-E_{ij})^2}{E_{ij}} = \sum_{i=1}^{k}\sum_{j=1}^{c}\frac{{x_{ij}}^2}{E_{ij}} - n$ n = sample size k=number of groups/treatments r=number of categories of classes for the variable	x_{ij}= observed number of cases for the category or class i under treatment j E_{ij}= number of cases expected for the category or class i under the treatment j
Common Hypothesis	*Ho : There is independence between the two groups and the variable of interest*	Ho assumes that the two groups/samples/ treatments are independent of the classes of the variable or that there is no interaction between the groups and the variable of interest
Small expected values	The chi-square test assumes a minimum of frequencies within each cell.	Usually a minimum of 5 subjects in each cell is expected.
Asymptotic Distribution/ Large Samples	X^2 *follows chi* − *square* (χ^2) with degree s of freedom df= (k-1)(r-1)	
Power	The test is usually the best where it is appropriate. It may not be appropriate for situations where order is taken into account	

NN. Extension of The Median Test

Extension of the Median Test help to assess whether k independent groups, with or without the same size, have likely been drawn from the same population or from populations with equal medians. The Median test is appropriate for ordinal scale measurements. There may also be specific instances where even interval measures may not have alternative to the Median Test, especially when observations are censored.

Observations may be classified as in the following contingency table:

Variable	Group or Sample			
	1	2	…	k
Observations above median	n_{11}	n_{11}	…	n_{11}
Observations below median	n_{11}	n_{11}	…	n_{11}

Statistic/ Concept	Formula	Comment
Test/ Variable Definition	$X^2 = \sum_{i=1}^{2}\sum_{j=1}^{k}\frac{(x_{ij}-E_{ij})^2}{E_{ij}} = \sum_{i=1}^{2}\sum_{j=1}^{k}\frac{{x_{ij}}^2}{E_{ij}} - n$ n = sample size k=number of groups/treatments	x_{ij}= observed number of cases for the category or class i under treatment j E_{ij}= number of cases expected for the category or class i under the treatment j
Common Hypothesis	$Ho:$ *There is independence among the k groups and the variale of interest*	
Asymptotic Distribution/ Large Samples	X^2 *follows chi − square* (χ^2) with degree s of freedom df= (k-1)	
Power	See Median Test	

OO. The Kruskal-Wallis One-Way Analysis of Variance by Ranks

The Kruskal–Wallis one-way analysis of variance by ranks, named after **American mathematical statistician William Kruskal** and **American Statistician-Economist W. Allen Wallis** is a non-parametric method for deciding whether multiple samples originated from different distributions.

The Kruskal-Wallis method tests the factual null hypothesis that the populations from which the samples originated have the same median. It is identical to a one-way analysis of variance with the data replaced by their ranks. It is also considered as an extension of the Mann–Whitney U test for three or more groups.

Since it is a non-parametric method, the Kruskal–Wallis test does not assume a normal population, but assumes that the variables under study have the same underlying continuous distributions. Hence, the test requires at least ordinal measurements.

To apply the test, the data are arranged into a two-way table with each column representing a sample, group or treatment and the rows being observations. All the observations from each group are combined and ranked as a series. Each observation is then replaced by its rank within the group. The Kruskall-Wallis statistic is then computed.

Observation	Sample or Treatment			
	1	2	…	k
1	X_{11}	X_{12}	…	X_{1k}
2	X_{21}	X_{22}	…	X_{2k}
3	X_{31}	X_{32}	…	X_{3k}
…	…	…	…	…
n	X_{b1}	X_{b2}	…	X_{bk}
Sum of Ranks	X_{*1}	X_{*2}	…	X_{*k}

Statistic/ Concept	Formula	Comment
Test/ Variable Definition	$KW = \left[\frac{12}{n(k+1)}\sum_{j=1}^{k} n_j \bar{r}._j{}^2\right] - 3n(k+1)$ k= number of treatments or groups (columns) n= number of observations within each group r_{*j}= Sum of the ranks in sample or group j $\bar{r}_{*j}$= Average of the ranks in sample or group j	The average ranks of the combined sample is $\bar{r} = (n+1)/2$
Common Hypothesis	$Ho\colon \theta_1 = \theta_2 = \cdots = \theta_k$ $H_1\colon \theta_i \neq \theta_j$, for at least two of the treatments	*There is no differences in the medians across all the treatments*
Asymptotic Distribution /Large Samples	KW is approximated by χ^2 with df=k-1	
Correction for Ties	$KW = \dfrac{\left[\frac{12}{n(k+1)}\sum_{j=1}^{k} n_j \bar{r}._j{}^2\right] - 3n(k+1)}{1 - \left[\sum_{i=1}^{g}(t_i{}^3 - t_i)\right]/(n^3 - n)}$	
Power-Efficiency	The power-efficiency of Kruskal-Wallis test may be compared to F test as n increase.	

PP. The Jonckheere Test for Ordered Alternatives

The ***Jockeree test for Ordered Alternatives*** is an extension of the Kruskal Wallis test. It tests whether the groups (or measures) are the same versus the alternative that the groups or measures are ordered in a specific a priori sequence. The a priori order of the groups is required for the test.

To apply the test, the data are arranged into a two-way table with each column representing a sample, group or treatment arranged in the a priori hypothetic order and the rows being observations. Consequently, the groups are ordered from the group 1 with the hypothetically lowest median to the group k with the largest median.

All the observations from each group are combined and ranked as a series. The **Jonckheere method** involves counting the number of times an observation in the group or sample *i* is preceded by the observation in the group or sample *j*. This count, also known as **the Mann-Whitney count** constitutes the basis for the test.

Statistic/ Concept	Formula	Comment
Test/ Variable Definition	$$U_{ij} = \sum_{h=1}^{n_i} \#\,(X_{hi}, \mathrm{j})$$ $\#\,(X_{hi}, \mathrm{j})$= number of times X_{hi} from group i precedes or is smaller than any observation in group j; with i<j. $$J = \sum_{i<j}^{k} U_{ij} = \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} U_{ij}$$	b= the number of blocks, groups or treatments n= the sample size.
Common Hypothesis	$Ho: \theta_1 = \theta_2 = \cdots = \theta_k$ $H_1: \theta_1 \leq \theta_2 \leq \cdots \leq \theta_k,$	*There is no differences in the medians across all the treatments*
Asymptotic Distribution/ Large Samples	$$\mu_J = \frac{n^2 - \sum_{j=1}^{k} n_j^2}{4}$$ $$\sigma^2{}_J = \frac{1}{72}\left[n^2(2n+3) - \sum_{j=1}^{k} n_j^2(2n_j+3)\right]$$ $$J^* = \frac{J - \mu_J}{\sigma_J}$$	J^* is approximately normally distributed with zero mean and standard deviation one.
Power-Efficiency	The power-efficiency of Jonckheere test is around 3/π when compared against an appropriate t or F test. It is asymptotically similar to Kruskall-Wallis	

QQ. The Cramer Coefficient C

The Cramer Coefficient C, named after the **Swiss mathematician Gabriel Cramer**, measures the degree of association or relation between two sets of attributes or variables, when the variables are measured at nominal scale. The Cramer Coefficient makes no assumptions regarding the continuity or the ordering of the variables or the attributes.

The coefficient is measured using a contingency table. The setting of the contingency table has no influence on the Cramer coefficient. The data may consist of any number of categories.

The **Cramer test** assumes that there are two sets of unordered categorical variables. If the variables are X and Y, the contingency table may is as follows:

Variable Y	Variable X				Total
	X1	X2	...	X_k	
Y1	X_{11}	X_{12}	...	X_{1k}	X_{1*}
Y2	X_{21}	X_{22}	...	X_{2k}	X_{2*}
Y3	X_{31}	X_{32}	...	X_{3k}	X_{3*}
...	...	...	...	...	...
Yr	X_{r1}	X_{r2}	...	X_{rk}	X_{b*}
Total	X_{*1}	X_{*2}	...	X_{*k}	n

Statistic/ Concept	Formula	Comment
Test/ Variable Definition	$C = \sqrt{\dfrac{X^2}{n(L-1)}}$ $X^2 = \sum_{i=1}^{r}\sum_{j=1}^{k}\frac{(x_{ij}-E_{ij})^2}{E_{ij}} = \sum_{i=1}^{r}\sum_{j=1}^{k}\frac{x_{ij}^2}{E_{ij}} - n$= number of treatments or groups	L= the minimum of the number of rows or columns in the contingency table. x_{ij}= observed frequency for cell(i,j) (attributes i of variable X and j of variables Y) E_{ij}= expected frequency for cell (i,j)
Common Hypothesis	*Ho: there is no association between X and Y*	H1 assumes that there is a relation/association among the two variables in the population
Asymptotic Distribution/ Large Samples	C is approximated by χ^2 with df=(r-1)(k-1)	
Power-Efficiency	Cramer Cis attractive for its relax assumptions but may be not powerful in some cases.	

RR. The Phi Coefficientr_ϕ for 2x2 Tables

The Phi Coefficientr_ϕ measures the extent of association or relation between two sets of variables measured on the nominal scale. The two variables may only take two values each. It is similar to Cramer's coefficient C. Every subject in each sample may fall in one or the other class based on the observation value.

Observations may be classified as in the following contingency table:

Variable X	Variable Y 0	1	Combined Frequency
0	A	B	A+B
1	C	D	C+D
Total	A+C	B+D	n

The test may be one or two-tailed.

Statistic/ Concept	Formula	Comment
Test/ Variable Definition	$r_\phi = \frac{\lvert AD-BC\rvert}{(A+B)(C+D)(A+C)(B+D)}$ $X^2 = \frac{n(\lvert AD-BC\rvert - \frac{n}{2})^2}{(A+B)(C+D)(A+C)(B+D)}$	A, B, C, D= number of subjects in one category of X and Y X^2 is distributed as Chi-square with df=1; .
Common Hypothesis	$Ho:$ *the variables are not related*	
Asymptotic Distribution/ Large Samples	The test may be estimated by a chi-square for large samples and Fisher for small samples.	
Power	The test is similar to Cramer's C	

SS. The Spearman Rank-Order Correlation Coefficient r_s

The ***Spearman's Rank-order correlation coefficient*** or ***Spearman's rho***, named after the **British psychologist Charles Spearman** and often denoted by the Greek letter **ρ (rho)** or as r_s, is a measure of statistical dependence or association between two variables.

The two variables must be measured at least at the ordinal scale so that subjects may be ranked into two ordered series. 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.

Statistic/ Concept	Formula	Comment
Test/ Variable Definition	$r_s = \frac{\sum x^2 + \sum y^2 - \sum d^2}{2\sqrt{\sum x^2 \sum y^2}}$ $r_s = 1 - \frac{6\sum {d_i}^2}{n^3 - n}$	$d_i = X_i - Y_i$, the difference of the ranks between the two variables n = sample size.
Common Hypothesis	$Ho:$ *the variables are not related*	
Asymptotic Distribution/ Large Samples	For large samples, z=$r_s\sqrt{n-1}$ is approximately normally distributed with mean 0 and standard deviation 1.	
Power-efficiency	The Spearman rank-order correlation coefficient has relatively high efficiency when compared to the most powerful parametric test, the Pearson product-moment correlation coefficient r = $\frac{\sum xy}{\sqrt{\sum x^2 \sum y^2}}$	

TT. The Kendall Rank-Order Correlation Coefficient τ

The **Kendall Rank-Order correlation coefficient**, commonly referred to as **Kendall's tau (τ)** coefficient and named after the **British statistician Sir Maurice George Kendall**, is a statistic used to assess the association between two measured quantities.

Specifically, it is a measure of rank correlation; that is, the similarity of the orderings of the data when ranked by each of the quantities.

Statistic/ Concept	Formula	Comment
Test/ Variable Definition	$\tau = \frac{\#\, agreements - \#\, disagreements}{total\ number\ of\ pairs}$ $\tau = \frac{2S}{n(n-1)}$ $r_s = 1 - \frac{6\sum {d_i}^2}{n^3 - n}$	n = sample size.
Common Hypothesis	$Ho:$ *the variables are not related*	
Asymptotic Distribution/ Large Samples	Mean $\mu_T = 0$ Variance: $\sigma_T^2 = \frac{2(2n+5)}{9n(n-1)}$ Z=$\frac{T-\mu_T}{\sigma_T}$ approximately follows normal distribution with mean 0 and standard deviation 1	
Power-efficiency	The Kendall τ and Spearman rank-order correlation coefficient have similar power-efficiency	

UU. The Kendall Partial Rank-Order Correlation Coefficient $T_{xy,z}$

The ***Kendall Partial Rank-Order*** correlation coefficient handles situations where a third variable Z is the source of associations between two variables X and Y.

The effects of variation due to the third variable are eliminated by keeping the variable constant while assessing the correlation between the two variables X and Y.

Specifically, Z scores are ordered, and the agreements between Z and the other variables are assessed.

X pair	Y pair: Sign agrees with the Z sign	Sign disagrees with the Z sign	Combined Frequency
Sign agrees with the Z sign	A	B	A+B
Sign disagrees with the Z sign	C	D	C+D
Total	A+C	B+D	n

Statistic/ Concept	Formula	Comment
Test/ Variable Definition	$T_{xy,z} = \frac{T_{xy} - T_{xz}T_{yz}}{\sqrt{(1 - T_{x,z}{}^2)(1 - T_{yz}{}^2)}}$	n = sample size.
Common Hypothesis	$Ho: \ the\ variables\ are\ not\ related$	
Asymptotic Distribution/ Large Samples	Mean $\mu_T = 0$ Variance: $\sigma^2_{Txy,z} = \frac{2(2n+5)}{9n(n-1)}$ Z=$\frac{3T_{xy,z}\sqrt{n(n-1)}}{\sqrt{2(2n+5)}}$ approximately follows normal distribution with mean 0 and standard deviation 1	
Power-efficiency	Not clear to assess	

VV. The Kendall Coefficient of Concordance W

Kendall's W, also known as Kendall's coefficient of concordance is a normalization of the statistic of **the Friedman test**, and can be used for assessing agreement or correlation among multiple independent measures.

Kendall's W ranges from 0 (no agreement) to 1 (complete agreement).

Suppose, for instance, that a number of raters have been asked to rank subjects, from most important to least important. **Kendall's W** can be calculated from these ranks:

- If the test statistic **W is 1**, then all the raters have been unanimous, and each rater has assigned the same order to the list of objects or subjects.
- If **W is 0**, then there is no overall trend of agreement among the raters, and their responses may be regarded as essentially random. Intermediate values of W indicate a greater or lesser degree of unanimity among the various responses.

While tests using the standard Pearson correlation coefficient assume normally distributed values and compare two sequences of outcomes at a time, Kendall's W makes no assumptions regarding the nature of the probability distribution and can handle any number of distinct outcomes.

W is linearly related to the mean value of the Spearman's rank correlation coefficients between all pairs of the rankings over which it is calculated.

Rater	Subject			
	1	2	…	n
1	R_{11}	R_{12}	…	R_{1k}
2	R_{21}	R_{22}	…	R_{2k}
3	R_{31}	R_{32}	…	R_{3k}
…		…	…	…
k	R_{k1}	R_{k2}	…	R_{bk}
Sum of Ranks	R_{*1}	R_{*2}	…	R_{*k}

Statistic/ Concept	Formula	Comment
Test/ Variable Definition	$average(r_s) = \frac{kW-1}{k-1}$ $W = \frac{\sum_{i=1}^{n}(\bar{R_i} - \bar{R})^2}{n(n^2-1)/12}$ $= \frac{12\sum_{i=1}^{n} {R_i}^2 - 3n(n+1)^2}{n(n^2-1)}$	k = number of sets of rankings, i.e., raters n=number of subjects $\bar{R_i}$ average rank for subject i $\bar{R}$ average or grand mean of all ranks
Common Hypothesis	Ho $:$ *the variables are not relate*	
Asymptotic Distribution/ Large Samples	$X^2 = k(n-1)W$ approximately follows chi-square distribution with df = n-1.	
Correction for Ties	$W = \frac{12\sum_{i=1}^{n} {R_i}^2 - 3k^2 n(n+1)^2}{k^2 n(n^2-1) - k\sum T_j}$, where the correction factor: $T_j = \sum_{i=1}^{g_j}({t_i}^3 - t_i)$, t_i is the size of the jth set of tied ranks in the ith group and g_i= the number of sets of tied ranks in the ith group	
Power-efficiency	Not clear to assess but efficiency increases with the sample size.	

WW. The Kendall Coefficient of Agreement u for Paired Comparisons or Rankings

Kendall's W tests the agreement or correlation among multiple independent measures. The concept is based on ratings, with a number of raters ranking subjects, from most important to least important.

In some experiments, raters are not asked to rank, but to indicate their **preferences for one of a pair of two objects presented at the same time**. Each object is usually paired with all other objects. Such experiment is known as a ***paired comparison***. The Kendall Coefficient of Agreement u is suited for such cases.

All the preferences from all the raters are combined in a preference matrix as follows:

Object	Raters/Judges			
	1	2	..	n
1	-	x_{12}		x_{1n}
2	x_{21}	x_{22}		x_{2n}
3	x_{31}	-		x_{3n}
...				
n	x_{k1}	x_{k2}		-

With n object, the possible number of pairs is: $\binom{n}{2} = (n)(n-1)/2$. With k raters, in case of complete agreement $(n)(n-1)/2$ cells will have frequencies equal to k and the remaining $= (n)(n-1)/2$ will have frequency equal to zero. If there is complete agreement, u will be equal to 1, whereas u=0 in case of complete lack of agreement.

Kendall's u has advantage over **Kendall's W** in the sense that it averages all the rankings for all the raters.

Statistic/ Concept	Formula	Comment
Test/ Variable Definition	$u = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}\binom{x_{ij}}{2}}{\binom{k}{2}\binom{n}{2}} - 1$ $= \frac{8\sum_{i=1}^{n} a_{ij}^2 - k\sum_{i=1}^{n} a_{ij}}{k(k-1)n(n-1)} + 1$	k = number of raters n=number of subjects x_{ij}: frequencies corresponding to the agreement between subject i and j
Common Hypothesis	$Ho:$ *there is no agreement among the raters*	
Asymptotic Distribution/ Large Samples	$X^2 = \binom{n}{2}[1 + u(k-1)]$ approximately follows chi-square distribution with df $= \binom{n}{2}$ = n(n-1)/2.	
Test with Ranks	$X^2 = \frac{6(2n+5)\binom{n}{2}\binom{k}{2}}{(k-2)(2n^2+6n+7)}\lvert u\rvert + f$ where f= $\frac{2(2n+5)^3\binom{n}{2}\binom{k}{2}}{(k-2)^2(2n^2+6n+7)^2}$ is the number of the degrees of freedom and needs to be rounded when not an integer.	
Power-efficiency	Kendall's u is relatively close to the Chi-square Goodness of Fit test.	

XX. The Correlation between Several Judges and a Criterion Ranking Tc

The Spearman rank-order correlation coefficient r_s and the Kendall rank-order correlation coefficient T assess the agreement between two rankers, by providing an index of correlation. Kendal coefficient of concordance W and Kendall coefficient of agreement u help to assess the agreement and concordance among the raters.

In some instances, experimenters are interested in agreement or concordance between the ratings and a specified criterion. The **Criterion Ranking test Tc** is suited for these instances where the focus on the correlation between k sets of rankings and a criterion ranking.

In the preference matrix, objects are listed in the order of the criterion ranking. Then, the frequencies are summed separately, for the upper triangle of the matrix as $\sum^{+} x_{ij}$on the one hand and the lower triangle on the other as $\sum^{-} x_{ij}$.

Object	Object			
	1	2	..	n
1	x_{11}	x_{12}		x_{1n}
2	x_{21}	x_{22}		x_{2n}
3	x_{31}	-		x_{3n}
…				
k	x_{k1}	x_{k2}		x_{kn}

Statistic/ Concept	Formula	Comment
Test/ Variable Definition	$T_c = \frac{2(\sum^{+} x_{ij} - \sum^{-} x_{ij})}{kn(n-1)} - 1$ $= \frac{4\sum^{+} x_{ij}}{kn(n-1)}$- 1 $= 1 - \frac{4\sum^{-} x_{ij}}{kn(n-1)}$-	k = number of raters n=number of subjects x_{ij}: frequencies corresponding to the agreement between subject i and j
Common Hypothesis	$Ho:$ *there is no agreement among the raters*	
Asymptotic Distribution/ Large Samples	To test the hypothesis Ho: Tc=0 vs H1: Tc>0; use the statistic: Z= $\left[T_{c\pm} \frac{2}{kn(n-1)}\right] \frac{3\sqrt{kn(n-1)}}{\sqrt{2(2n+5)}}$ which is approximately distributed with mean zero and standard deviation one.	

YY. The Cohen's Kappa Statistic κ

The Cohen's Kappa statistic, named after the **US psychologist John Cohen**, tests the agreement or concordances among the raters when the measurements are nominal (categorical).

Specifically, the test is a measure of inter-rater agreement for categorical items. The ranking matrix may look like as follows:

Object	Category				
	1	2	...	k	Total
1	x_{11}	x_{12}	...	x_{1k}	x_{1*}
2	x_{21}	x_{22}	...	x_{2k}	x_{2*}
3	x_{31}	...	...	x_{3k}	x_{3*}
...			...	...	
n	X_{n1}	X_{n2}		X_{nk}	x_{n*}
Total	x_{*1}	x_{*2}		x_{*k}	

Each of the **n objects** is assigned to one of the **m categories** by each of **k raters.** Xij is the number of the raters that assign object i to category j. If x*j is the number of ratings that assign the objects to category j.

If the raters are in agreement on a category for a given object, the frequency for the object in the category is k, other categories having 0 frequency. The kappa coefficient of agreement is the ratio of the proportion of times that the raters agree to the maximum proportion of times.

Statistic/ Concept	Formula	Comment
Test/ Variable Definition	$\kappa = \frac{P(A) - P(E)}{1 - P(E)}$ $P(E) = \sum_{j=1}^{m} p_j^2 = \sum_{j=1}^{m} \left(\frac{C_j}{nk}\right)^2$ $P(A) = \frac{1}{n}\sum_{j=1}^{m} X_{i*} = \left[\frac{1}{nk(k-1)} \sum_{i=1}^{n} \sum_{j=1}^{m} x_{ij}^2\right]$	k = number of raters n=number of subjects x_{ij}: frequencies corresponding to the number of raters that assign object i in category j P(A)= proportion of times that k raters agree P(E) = proportion of times that the k raters may agree du chance.
Common Hypothesis	$Ho: \text{ there is no agreement among the raters}$	
Asymptotic Distribution / Large Samples	To test Ho: κ=0 against H1:κ>0; Var(κ)= $\left[\frac{2}{nk(k-1)}\right]\left[\frac{P(E)-(2k-3)[P(E)]^{2}+2k(k-2)\sum_{j=1}^{m} p_j^3}{}\right]$ Z=$\kappa/\sqrt{var(\kappa)}$ is approximately distributed with mean zero and standard deviation one.	
Power-efficiency	Usually more robust measure than simple percent agreement calculation	

ZZ. The Gamma Statistic Γ

The ***Gamma statistic*** tests the agreement or concordances between two variables measured at the ordinal scale. The approach in applying the **Gamma Statistic** is similar to **Kandall's tau**.

The categories for the two variables may consist of any number. The contingency table may look as follows:

Ordinal Variable B	Ordinal Variable A				
	A1	A2	…	Ak	Total
B1	x_{11}	x_{12}		x_{1k}	x_{1*}
B2	x_{21}	x_{22}		x_{2k}	x_{2*}
B3	x_{31}	-		x_{3k}	x_{3*}
…					
Br	X_{r1}	X_{r2}		-	x_{n*}
Total	x_{*1}	x_{*2}		x_{*k}	n

Statistic/ Concept	Formula	Comment
Test/ Variable Definition	$\Gamma = \frac{\#\,agreements - \#\,disagreements}{\#\,agreements + \#\,disagreements}$ $\#agreements = \sum_{i=1}^{r-1}\sum_{j=1}^{k-1} x_{ij} \sum_{p=i+1}^{r} \sum_{q=j+1}^{k} x_{pq}$ $\#disagreements = \sum_{i=1}^{r-1}\sum_{j=2}^{k} x_{ij} \sum_{p=i+1}^{r} \sum_{q=1}^{j-1} x_{pq}$	n = sample size.
Common Hypothesis	$Ho:\ the\ variables\ are\ independent$	
Asymptotic Distribution/ Large Samples	Mean: $\gamma = 0$ Variance: $\sigma_\Gamma^2 \leq \frac{n(1-G^2)}{\#agreements+\#disagreements}$ $Z = \frac{\Gamma-\gamma}{\sigma_\Gamma}\sqrt{\frac{\#agreements+\#disagreements}{n(1-G^2)}}$ approximately follows normal distribution with mean 0 and standard deviation 1	

AAA. The Lambda Statistic λ_b

The ***Lambda statistic*** λ_b developed by the **American mathematical statisticians William Henry Kruskal** and **Leo A. Goodman** is a measure of proportional reduction in error in cross tabulation analysis of two variables.

The test makes a few assumptions regarding the scale of the variables, except the fact that the variables are nominal, but not ordered. For any sample with a nominal independent variable and dependent variable (or ones that can be treated nominally), it indicates the extent to which one variable can be predicted by another known variable.

The contingency table may look like as follows:

Nominal Variable B	Nominal Variable A				
	A1	A2	…	Am	Total
B1	x_{11}	x_{12}		x_{1k}	x_{1*}
B2	x_{21}	x_{22}		x_{2k}	x_{2*}
B3	x_{31}	…		x_{3k}	x_{3*}
…					
Br	X_{r1}	X_{r2}		…	x_{n*}
Total	x_{*1}	x_{*2}		x_{*k}	n

Hence, the test is a measure of the reduction in the error in predicting one variable by another known variable

Statistic/ Concept	Formula	Comment
Test/ Variable Definition	$\lambda_b = \frac{\sum_{j=1}^{k} \max(x_{*j}) - \max(x_{i*})}{n - \max(x_{i*})}$ $\max(x_{*j})$: largest frequency in column j $\max(x_{i*})$: largest frequency in row i	n = sample size.
Common Hypothesis	$Ho:$ *the variables are independent*	
Asymptotic Distribution/ Large Samples	Mean: λ_b Variance: $\sigma_\lambda^2 = \frac{\left(n - \sum_{j=1}^{k} \max(x_{*j})\right)\left(\sum_{j=1}^{k} \max(x_{*j}) + \max(x_{i*}) - 2\Sigma' \max(x_{*j})\right)}{[n - \max(x_{i*})]^3}$ $Z = \frac{\Gamma - \lambda_b}{\sigma_\lambda}$ approximately follows normal distribution with mean 0 and standard deviation 1	

BBB. Asymmetrical Association For Ordered Variables: Somers's d_{BA}

The ***Somers's d_{BA} statistic***, named after the **American sociologist Robert H. Somers** is an asymmetric measure of association between two variables.

Given a predictor variable A and an outcome variable B, d_{BA} may be estimated as a measure of the effect of B on A , or as a performance indicator of B as a predictor of A.

To compute the statistic d_{BA} determine the frequencies above to the left, M_{ij}^{-} and above to the right, $M_{ij}^{\mp}$ of the cell (i,j) and the frequencies below to the left, N_{ij}^{-} and below to the right, $N_{ij}^{\mp}$ of the cell (i,j).

$$M_{ij}^{-} = \sum_{p=1}^{i-1} \sum_{q=1}^{j-1} x_{pq}$$

$$M_{ij}^{+} = \sum_{p=1}^{i-1} \sum_{q=j+1}^{k} x_{pq}$$

The contingency table may look like as follows:

Nominal Variable B	Nominal Variable A			
	A1	A2	Am	Total
B1	x_{11}	x_{12}	x_{1k}	x_{1*}
B2	x_{21}	x_{22}	x_{2k}	x_{2*}
B3	x_{31}	-	x_{3k}	x_{3*}
...				
Br	X_{r1}	X_{r2}	-	x_{n*}
Total	x_{*1}	x_{*2}	x_{*k}	n

Statistic/ Concept	Formula	Comment
Test/ Variable Definition	$d_{BA} = \frac{\#\,agreements - \#\,disagreements}{\#pairs\ not\ tied\ on\ variable\ A} = \frac{\#\,agreements - \#\,disagreements}{n^2 - \sum_{i=1}^{r} x_{i*}^2}$	n = sample size.
Common Hypothesis	$Ho:\ the\ variables\ are\ independent$	
Asymptotic Distribution/ Large Samples	Mean: $d_{BA} = 0$ Variance: $\sigma_{\Gamma}^2 = \frac{4\sum_{i=1}^{r}\sum_{j=1}^{n} n_{ij}\ (N_{ij}^{+} + M_{ij}^{+} - N_{ij}^{-} - M_{ij}^{-})^2}{\left[N^2 - \sum_{j=1}^{k} x_{*j}^2\right]^2}$ Z= $\frac{d_{BA}}{\sqrt{var(d_{BA})}}$ approximately follows normal distribution with mean 0 and standard deviation 1	

VIII. Other Key Tests

A. Ansari-Bradley Test

Some research problems may need to test the spreads for two populations with assumed equal central tendency. In the parametric case, the F-Test for equality of variances will be used.

In the non-parametric situations, the **Ansari-Bradley test** is used to test the differences in spread between two populations when the population medians are assumed equal and the samples from two populations are independent. The measures are assumed continuous.

Ansari-Bradley scores are similar to Siegel–Tukey scores, with the difference that Ansari-Bradley assigns the same scores to extreme ranks.

Approach:

- **Step1: Center** around the median the observations within each of the two samples. That is if X = {X1, X2, … Xn1} is the set of observations in Sample 1 with median $\mu 1$ and Y={Y1, Y2, … Yn2) is the set of observations in Sample 2, with median $\mu 2$, then centered observations are {X1-μ1, X2-μ1, ….Xn1 -μ1} and {Y1-μ2, Y2-μ2, ….Yn2 -μ2}.
- **Step 2:** Combine the two location (median)-adjusted samples and list in increasing order. Assume n= n1 + n2.
  Step 3: Assign the score (rank) i to the I th and (n + 1 − i)th ordered values. In case of ties, assign average scores. Thus, scores increase from both ends towards the center of the ordered sample.
  Ansari-Bradley scores are computed as follows, with R_j, the rank of observation j, and $a(R_j)$, the score of observation j:

$$a(R_j) = \frac{n+1}{2} - \left|R_j - \frac{n+1}{2}\right|,$$

- **Step 4:** Ansari-Bradley statistic W is the sum of the scores(ranks).

- **Step 5:** Look up the critical value in Ansari-Bradley tables in n<20, otherwise use normal approximation.
- From the table, given the significance level c, the upper critical value of W is the smallest x in the table such that P(W ≥ x) ≤ c and the lower critical value of W is x-1, such that P(W ≥x)≥1−c.

For large samples, approximate the standard normal distribution Z with W*, and use the normal tables to fin p-value.

If n=n1+n2 is even:

$$W^* = \frac{W - [n_1(n_1+n_2+2)/4]}{\sqrt{n_1(n_1+n_2+2)n_2(n_1+n_2-2)/[48(n_1+n_2-1)]}}$$

If n=n1+n2 is odd:

$$W^* = \frac{W - [\frac{n_1(n_1+n_2+2)^2}{4(n_1+n_2)}]}{\sqrt{n_1(n_1+n_2+2)n_2(3+(n_1+n_2)^2/[48(n_1+n_2)^2]}}$$

B. Conover's Squared Ranks Test for Equality of Variance

The assumptions of the **Conover Test also known as Conover Squared Ranks Test for Variances,** are similar to those for Ansari-Bradley test. It tests whether the samples come from populations with equal dispersion.

The Conover Test uses Conover scores calculated as the squared ranks of the absolute deviations from the sample means. For each observation i, the absolute deviation from the mean is computed as:

$$U_i = \left|X_{i(j)} - \bar{X}_j\right|$$

where $X_{i(j)}$ is the value of observation i , within sample j and $\bar{X}_j$ is the mean of sample j.

The values U_i are then ranked, assigning average ranks to ties.

The Conover score, $Score_j = (Rank(U_i))^2$

C. Normal Scores

Normal scores, also known as Van der Waerden are the quantiles of a standard normal distribution. Hence they are also known as quantile normal scores. These scores are powerful for normal distributions.

Van der Waerden Normal scores are computed as:

$$a(R_j) = \Phi^{-1}\left(\frac{R_j}{n+1}\right)$$

where Φ is the cumulative distribution function of a standard normal distribution.

D. Mood Scores

Mood scores are the square of the difference between the observation rank and the average rank. Mood scores are computed as follows:

$$a(R_j) = \left(R_j - \frac{n+1}{2}\right)^2$$

E. Klotz Scores

Klotz Scores are the squares of the Normal or **Van der Waerden** scores.

$$a(R_j) = \left(\Phi^{-1}\left(\frac{R_j}{n+1}\right)\right)^2$$

IX. Non Parametric Survival Tests: Tarone-Ware, Log-Rank and Gehan-Breslow Test

The Log-Rank test, also known as **the Mantel log-rank test**, **Cox Mantel log-rank** test named after the English statistician, Sir David Roxbee Cox or the **Mantel- Haenszel** test, named after the American statisticians Nathan Mantel and William Haenszel, is commonly used in testing and comparing two survival distributions.
It is applicable to data where there is progressive censoring. The test assigns equal weights to early and late failure. Under the null hypothesis, the hazard functions for the two groups are parallel.

Tarone-Ware and **Gehan-Breslow** tests, named after the American and English statisticians Edmund A. Gehan, Norman E. Breslow are used to compare two survival distribution or functions.
They are versions of the log ranks, after assigning the weight to the observed minus the expected score at any given time,

- The **Log-Rank** uses the **weight of 1.0** for each observation.
- **Gehan-Breslow,** sometimes known as **Wilcoxon** uses the weight corresponding to the number of observations at risk. r_j .
- The **Tarone-Ware** test uses weights corresponding to the square root of the number of observations with the lower rank or observations at risk, $\sqrt{r_j}$
- The **Cox Mantel** is also similar to the log-rank test
- The Peto-Peto modification of the Gehan-Wilcoxon test, named after the British statisticians Sir Richard Peto and Julian Peto is similar to Breslow's test. It is more suited for situations where the hazard ratio between the groups is not constant. It is considered more powerful than Gehan-Breslow's test when the sample sizes are small.

The null hypothesis is assumes that the populations have the same survival distributions, whereas the alternative assumes different survival distributions

1. Log-Rank Test

- The test compares two 2 groups, but may be extended to more groups
- Approach: Find the expected number of failures in group 2 and compare with the observed number of failures.

$$X = \frac{(E_2 - O_2)}{Var(E_2 - O_2)}^2 \ follows\ a\ chi - square\ with\ 1\ degree\ of\ freedom\ (\chi_1^2)$$

where $\frac{r_{1j} r_{0j} d_j (r_j - d_j)}{r_j^2 (r_j - 1)}$ is the variance and $E_2 = \frac{r_{1j} d_j}{r_j}$,

with r_{1j} = number of observations at risk at time j in group 1;

r_{0j} = number of observations at risk at time j in group 0;

d_j = combined number of failures/events at time j;

r_j = combined number of observations at risk at time j;

and O_2= total number of failures in group 2, E_2 = expected number failures in group 2.

For extension to more than two groups, g ≥ 2 groups possible, the covariance of the total number of failures and the expected number of failures is used:

$$X = \sum \frac{(E_i - O_i)}{Var(E_i - O_i)}^2 \; follows\ a\ \ chi - square\ with\ g - 1\ degree\ of\ freedom\ (\chi_g^2)$$

2. Stratified Log-Rank Test

.

- The test is an extension of the Log-Rank Test
- Allows controlling for an additional variable, known as strata, hence “stratified”
- Approach: Split data into strata, depending on value of stratified variable •
- Calculate $E_2 - O_2$ scores within strata
- Sum $E_2 - O_2$ cross strata.

3. Tarone-Ware, Gehan-Breslow or Generalized Wilcoxon Test

- The test is an extension of the Log-Rank Test
- Allows weights on observations with a Weights variable, while controlling or not for strata
- It is more powerful than the log-rank test when the hazard functions are not parallel and in case of little censoring. When censoring is pervasive, it has low power when censoring. It gives more weight to early failures.
-

$$\chi_{tw}^2 = \frac{\left[\sum_{j=1}^{K} w_i \left(d_{1j} - r_{1j} * d_j / r_j\right)\right]^2}{\sum_{j=1}^{K} \frac{w_j^2 r_{1j} r_{0j} d_j (r_j - d_j)}{r_j^2 (r_j - 1)}}$$