Kolmogorov Smirnov test (K.S.Test)

A manufacturer of readymade garment Conducts a market Survey to Know the choice of brands A, B, C and D of 100 prospective Customers. The results are $A = \overset{O.F}{20}$, $B = 30$, $C = 18$, $D = 32$. Use Ks test at 5% l.o.s to Know

If Customers have any distinct brand preference.

The table Value of D for N larger than 35 at 5% l.o.s is $\frac{1.36}{\sqrt{N}}$

H_0: The Customers do not Show any distinct brand preference.

$\|F_o - F_e\|$	Brand	Observed Frequency	Cumu Obs. Fr	Cum Ob. Fr Prop (F_o)	Exd Frequ	Cu Fre F_e
0.05	A	20	20	0.2 $\frac{20}{100}$	25	25
0	B	30	50	0.5	25	50
0.07	C	18	48	0.48	25	75
0	D	32	100	1	25	100

K.S statistic is maximum of D is 0.07

The Value of D is $\frac{1.36}{\sqrt{N}} = \frac{1.36}{10} = 0.136$

2. From the following information

10m

(i) Calculate K.S statistic.

(ii) Can we Conclude that this distribution thus infact follow a normal Distribution You are given the tabulated Value of D_n for $n = 5$ at 10% L.O.S as 0.510

Test score :	51-60	61-70	71-80	81-90	91-100
observed fre :	30	100	440	500	130
Expected fr :	40	170	500	390	100

H_0 :

T.S	O.F	C.O.F	Expressed C.O.F (P) (F_0)	E.F	C.E.F	C.E.F (P) Fe	$\|F_0 - Fe\|$
51-60	30	30 $\frac{30}{1200}$	0.025	40	40 $\frac{40}{1200}$	0.0333	0.0083
61-70	100	130 $\frac{130}{1200}$	0.1083	170	210 $\frac{210}{1200}$	0.175	0.0667
71-80	440	570 $\frac{570}{1200}$	0.475	500	710	0.5917	0.1167
81-90	500	1070	0.8917	390	1100	0.9167	0.025
91-100	130	1200	1	100	1200	1	0

K.S statistic maximum of D_n is 0.1167

The Value of D is .

Non- Parametric test

A non-parametric test (sometimes called a distribution free test) does not assume anything about the Underlying distribution (for example that the data Comes from a normal distribution) Use non parametric test only if you have to (you know that assumptions like normality are being violated)

when to Use non-parametric test

Non- parametric tests are Used when your data isn't normal.

Therefore the key is to figure Out if you have normally distributed data.

For example, You Could look at the distribution of your data. If your data is approximately normal, then you can Use parametric Statistical tests.

Parametric test

Parametric tests assume Underlying statistical distributions in the data. Therefore, several Conditions of Validity must be met so that the result of a Parametric test is reliable. Non parametric

test do not rely on any distribution. Th
can thus be applied even if parametric Conditions of Validity are not met

Parametric Test	Non-parametric
Paired t-test	Wilcoxon Rank Sum test
Unpaired t-test	Mann Whitney test
Pearson Correlation	Spearman Correlation
One way Analysis of Variance	Kruskal Walli's Test

Parametric For large Sample	Non-parametric For Small Sample
Robust p-Value is approximately Correct even if population is not normally distributed	Not Robust If the population is not normally distributed the p-Value may be misleading
Powerful If the population is normally distributed the	If the population is normally distributed

p-Value will be approximately identical to the p-Value Obtained from a Parametric test	Will be higher than the p-Value obtained from a t-test Less powerful

Reasons to run non-parametric test

1) One or more assumptions of a Parametric test have been violated.

2) Your Sample Size is too Small to run a parametric test.

3) Your data has outlier that Cannot be removed.

4) You Want to test the median rather than the mean(you might want to do this if you have a very skewed distribution)

Main non-parametric test are.

(I) 1-Sample Sign test: Use this test to estimate the median of a population and Compare it to a reference Value or target Value.

II) 1-Sample Wilcoxon Signed rank test
With this test, you also estimate the population median and Compare it to a

reference / target Value. However, the test assumes your data Comes from a Symmetric distribution or Uniform distribution

Test for Randomness:

Randomness tests, in data evaluations are Used to analyze the distribution of a Set of data to See if it can be described as random

How do you test for randomness?

In the Stock market, run test of randomness is applied to know if the Stock price of a particular Company is behaving randomly, or if there is any pattern

Run test of randomness is basically based on the run Run is basically a Sequence of one symbol such as + or -

When testing for randomness we can use?

Run test of Randomess Running a Test of Randomness is a non-parametric method that is Used in Cases when the parametric test is not in Use In this test, two different random Samples from different population with different Continuous Cumulative distribution functions are obtained.

what is a measure of randomness?

Many articles and books write that entropy is the measure of randomness or disorder of the system. They say when a gas system is let expand the randomness increases etc. But they end up saying dS is the measure of increase in randomness and is called the entropy.

True randomness

Randomness is the lack of pattern or Predictability in events. Individual random events are by definition Unpredictable. but in many Cases the frequency of different Outcomes over a large number of events (or "trials") is predictable

what is randomness bias?

People are often biased in their perception of randomness in that they tend to see patterns in random distributions. This is a serious problem because the accurate perception of randomness Can be important in decision making.

Run - Test

A simple statistical test of the random-walk theory is a runs test. For daily data, a run is defined as a sequence of days in which the stock price changes in the same direction.

For Example: Consider the following Combination of Upward and downward price changes:

+ + - - + - + - - - + +.

A + sign means that the stock price increased, and a - sign means that the stock price decreased. Thus the example has 7 runs in 12 observations

Application:

Run test Can be Used to test.

The randomness of a distribution by taking the data in the given

Observations are not normally distributed. However, they should follow the same shape (i.e) both are bell shaped and skewed left.

Rank Correlation:

Spearman's Rank Correlation

Spearman's rank Correlation Co-efficient is a technique which can be Used to Summarise the strength and direction (negative or positive) of a relationship between two Variables. The result will always be between 1 and minus 1. Rank the two data Sets. The mean rank in this case is Calculated as $(5+6+7) \div 3 = 6$

$$P = 1 - \left\{ \frac{6 \leq d_i^2}{n(n^2-1)} \right\}$$

P = Spearman's rank Correlation

d_i = difference between the ranks of Corresponding Variables

n = number of observations

Sign test

The sign test is a statistical test to Compare the sizes of two groups. It is a non-parametric or "distribution free" test, which means the test doesn't assume the data Comes from a particular distribution, like the normal distribution. The sign test is an alternative to a one sample t-test or paired t test.

The null hypothesis for the sign test is that the difference between median is zero.

For a One Sample Sign test, where the median for a single sample is analyzed, see: One Sample Median test

How to Calculate a Matched sample Sign test.

The data should be from two samples The two dependent samples should be paired or matched. For example, depression scores from before a medical procedure and after

Non-parametric rank Correlation

There are two accepted measures of non-parametric rank Correlation.

i) Kendall and Spearman's rank Correlation Co-efficient.

ii) Correlation analysis measure the strength of the relationship between minus One and plus one.

Comparison of two population

A point estimate for the difference in two population means is simply the difference in the Corresponding sample means. The same five step procedure Used to test hypothesis Concerning a single population mean is Used to test hypothesis Concerning the difference between two population means

Median test

The Median test is a non-parametric test that is Used to test whether two (or more) independent groups differ in Central tendency - Specifically whether the groups have been drawn from a population with the same median. The null hypothesis is that the groups are drawn from populations with the same median

$$\text{Median} = l + \frac{N/2 - c}{f} \times i$$

l = Lowerlimit of Median class

f = frequency of Median class

i = class Interval.

c = preceeding Cumulative frequency from median class

If the total number of numbers (n) is an odd number,

$$\text{Median} = \left(\frac{n+1}{n}\right)^{th} \text{term}$$

If the total number of numbers (n) is an even number,

$$\text{Median} = \frac{\left(\frac{n}{2}\right)^{th} \text{term} + \left(\frac{n}{2}+1\right)^{th} \text{term}}{2}$$

Mean-whitney U test

The Mann-whitney U test is a non-parametric test that can be used in place of an unpaired t-test. The Mann whitney test is based on a comparison of every observation x_i in the first sample with every observation y_j in the other sample. The total number of pairwise comparisons that can be made is $n_x n_y$.

Assumption for the Mann whitney U test

1) The dependent variable should be measured on an ordinal scale or a continuous scale

2) The independent variable should be two independent categorical groups

3) Observations should be independent

In other words there should be no relationship between the two groups or within each group.

4) Observations are not normally distributed. However they should follow the same shape (ie both are bell-shaped and skewed left).

Mann whitney U test Formula

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1$$

$$U_2 = n_1 n_2 - U_1$$

where,

n_1 = Sample size of group 1

n_2 = Sample size of group 2

R_1 = Sum of ranks of group 1.

Wilcoxon Signed Rank test

The Wilcoxon Signed rank test is a non-parametric statistical hypothesis test Used to Compare two related Samples, matched samples, or repeated measurements on a single sample to asses whether their population mean ranks differ (i.e it is a paried difference test).

Two slightly different Versions of the test exist

* The Wilcoxon Signed rank test Compares your Sample median against a hypothetical median.

* The Wilococxon matched pairs Signed rank test Computes the difference between each set of matched pairs, then follows the same procedure as the signed rank test to Compare the Sample against some median

The term "Wilcoxon" is often Used for either test. This Usually isn't Confusing, as it should be obvious if the data is matched or not matched

The "null Hypothesis" for this test is that the median of two Samples are equal. It is generally Used.

(i) As a non-parametric alternative to the one Sample t test or paired t test

(ii) For ordered (ranked) Categorical Variables Without a numerical Scale

Wilcoxin Rank Sum test

$$Z = \frac{W_1 - W_e}{\sigma_W} = \frac{W_1 - W_e}{\sqrt{n_1 n_2 (n_1 + n_2 - 1)/12}}.$$

Obtaining the expected rank to Compute the Z score

$$W_e = \frac{n_1 (n_1 + n_2 + 1)}{2}$$

Median test for Several Samples..

Median Test (for two Samples) The Median' Test, essentially a two sample Version of the Sign Test, is Used to determine whether the median of two independent Samples are equal. To perform this test you need to execute the following steps: Calculate the total median m of the Combination of the two Samples).

How to find the median of two Samples

To find the median, the data should be arranged in Order from least to greatest. If there is an even number of items in the data set, then the median is found by taking the mean (average) of the two middle most numbers

Kruskal Wallis test

The Kruskal Wallis test is the non-parametric alternative to the one way ANOVA. Non-parametric means that the test doesn't assume your data comes from a particular distribution. The H-test is used when the assumptions for ANOVA aren't met (like the assumption of normality). It is sometimes called the one-way ANOVA ranks as the ranks of the data values are used in the test rather than the actual data Points.

The test statistic used in this test is called H- Statistic. The hypothesis of the test are.

H_0: population medians are equal

H_1: population medians are not equal

The Kruskal Wallis test will tell you if there is a significant difference between groups. However it won't tell you which groups are different. For that you'll need to run a Post Hoc test

Assumptions for the Kruskal Wallis Test

Your Variables should have.

1. One independent Variable with two or more levels (independent groups). The test is more Commonly Used when you have three or more levels. For two levels Consider Using the mann-whitney u Test instead.
2. Ordinal Scale, Ratio Scale or Interval Scale dependent Variables
3. Your observations should be independent In otherwords there should be no relationship between the members in each group or between groups. For more information on this point See: Assumption of independence.
4. All groups should have the same shape distributions. Most software (SPSS, Minitab) will test for this Condition as Part of the test

Kruskal Walli's test Formula:

1. $$H = \frac{12}{N(N+1)} \sum_{i=1}^{g} n_i \left(\bar{r}_i - \frac{N+1}{2} \right)^2$$

2. Chi Square with $k-1$ degrees of freedom

$$W = \left[\frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} \right] - 3(n+1)$$

where,

n = Sum of Sample size in all groups

k = Number of Samples

R_i = Sum of ranks in i^{th} group

n_i = Size of i^{th} group.

Friedman's Test

Friedman's Test is a non-parametric test for finding differences in treatments across multiple attempts. Non-parametric means the test doesn't assume your data Comes from a particular distribution (like the normal distribution) Basically its Used in place of the ANOVA Test when you don't know the distribution of your data.

Friedman's test is an extension of Sign test Used when there are multiple treatments. In fact if there are only two treatments the two are identical.

Two Way Analysis of Variance by Ranks

Running the test.

Your data should meet the following requirements.

(I) Data should be Ordinal (e.g the Likert Scale or Continuous,

(II) Data Comes from a single group, measured on at least three different Occasions.

(III) The Sample Was Created with a random Sampling method.

(IV) Blocks are mutually independent (ie all of the pairs of are independent - One doesn't affect the other).

(V) Observations are ranked Within blocks with no ties.

The Null Hypothesis for the test is that the treatments all have identical effects or that the Samples differ in Some way. For example. they have different Centers. spread or shapes. The alternative hypothesis is that the treatments do have different effects

Friedman Formula

$$FM = \left[\frac{12}{(N*k*(k+1)}\right]*\Sigma R^2-[3*N*(k+1)]$$

where

n: number of Subjects

k - Number of treatments

R - The total ranks for each of the three Columns.

Kolmogorov - Smirnov Test

The Kolmogorov-Smirnov Goodness of fit Test (k- s test) Compares your data with a known distribution and lets you know if they have the same distribution. Although the test is non-parametric - it doesn't assume any Particular underlying distribution - It is Commonly Used as a test for normality to see if your data is normally distributed. It also Used to check the assumption of normality in analysis of

Variance.

More specifically the test compares a known hypothetical probability distribution (eg the normal distribution) to the distribution generated by your data - the empirical distribution function.

Calculating the Test Statistic

The K-s test statistic measures the largest distance between the EDF $F_{data}(x)$ and the theoretical function $F_0(X)$. measured in a Vertical direction (Kolmogorov as citied in Stephens 1992)

The test statistic is given by

$$D = \sup_{x} |F_0(x) - F_{data}(x)|$$

Where (for a two tailed test)

$F_0(X)$ = the cdf of the hypothesized distribution

$F_{data}(x)$ = the empirical distribution function of your observed data.

If D is greater than the Critical Value, the null hypothesis is rejected. Critical Values for D are found in

K-stest P-Value Test

Advantages and Dis advantages

Advant ages include:

* The test is distribution free. That means you don't have to know the Underlying population distribution for your data before running this test.
* The D Statistic (not tobe Confused with Cohen's D) Used for the test is easy to Calculate.
* It Can be Used as a goodness of fit test following regression analysis.
* There are no restrictions on Sample Size; Small Samples are acceptable.
* Tables are readily available

Although the K-S test has many advantages it also has a few limitations:

1. In Order for the test to work, you must Specify the location, Scale and Shape parameters. If these parameters are estimated from the data, it invalidates the test. If you don't know these parameters, you may Want to run a less formal test (like the One

Out lined in the empirical distribution function article).

2. It generally can't be Used for discoete distributions, especially if you are Using software (most software packages don't have the necessary extensions for discoete k-s Jest and the manual calculations are Convoluted).

3. Sensitivity is higher at the Center of the distribution and lower at the tails.

Test statistic.

$$Q = \frac{12}{mk(k+1)} \sum_{j=1}^{k} R_j^2 - 3m(k+1)$$

Where R_j^2 is the square of the rank total for group j $(j = 1, 2 .. k)$.

m is the number of independent blocks.

k is the number of groups or treatment levels

Kolmogorov Smirnov Test formula

$$KS = \max_j \sqrt{\frac{1}{n} \sum_i n_i \left[F(x_j) - F(x_j) \right]^2}$$

where $j = 1, 2 .. n$

The probability density of the solutions satisfies the forward Kolmogorov equation

$$\begin{cases} \dfrac{\partial P(t,x)}{\partial t} = -\dfrac{\partial(ap(t,x))}{\partial x} + \dfrac{b^2}{2}\dfrac{\partial^2(P(t,x))}{\partial^2 x} \\ \end{cases}$$

$$P(0,x) = \delta(x-x_0)$$

The solution of this PDE equation is

$$P(t,x) = \frac{1}{\sqrt{2\pi b^2 t}} \exp\left\{-\frac{(x-at-x_0)^2}{2b^2 t}\right\}$$

Chi-Square Formula:

$$\chi^2 = \frac{(x_1-\mu)^2}{\sigma^2} + \frac{(x_2-\mu)^2}{\sigma^2} + \dots + \frac{(x_k-\mu)^2}{\sigma^2}$$

$$= \sum_{i=1}^{k} \frac{(x_i-\mu)^2}{\sigma^2}$$

(or)

$$\chi^2 = \sum \frac{(O_i-E_i)^2}{E_i}$$

O_i = observed frequency

E_i = Expected frequency

Σ = the "sum of"

Main non-parametric test:

Friedman test: This test is Used to test for differences between groups with Ordinal dependent Variables. It Can also be Used for Continuous data if the one-Way ANOVA with repeated measures is inappropriate (i.e) some assumption has been Violated)

Goodman Kruska's Gamma: a test of association for ranked Variables.

Kruskall-Wallis test: Use this test instead of a one way ANOVA to find Out if two or more medians are different. Ranks of the data points are Used for the Calculations, rather than the data points themselves.

Man-Kendall Trend test - looks for trend in time Series data.

Mann-Whitney Test: Use this test to Compare difference between two independent groups when dependent Variables are either Ordinal or Continuous.

Mood's Median Test: Use this test instead of the sign test when you have two independent samples

Spearman Rank Correlation: Use when you want to find a Correlation between two sets of data.

Using the Normal Approximation with Wilcoxin Signed Ranks:

If the number of observations/ Pairs $n(n+1)/2$ is greater than 20, you can Use a normal approximation. this set of data meets this requirement ($12(12+1)/2 = 78$ There are several notifications/ Considerations for the Z-Score formula.

Use the smaller of W^+ or W^- for the test statistic

Use the following formula for the mean $\mu : n(n+1)/4$

Use the following formula for

$\sigma : \sqrt{n(n+1)(2n+1)/24}$

If you have tied ranks, you must reduce σ by $t^3 - t/48$ for each of t tied ranks. There are two tied ranks. $(5.5 + 5.5)$, so indicated by the sign of the coefficient : a + sign indicates a positive relationship.

a - sign indicates a negative relationship

We measure four types of Correlation

Pearson Correlation

Kendall rank Correlation

Spearman Correlation

the point - Biserial Correlation.

Pearson r- Correlation: pearson r Correlation is the most widely Used Correlation statistic to measure the degree of the relationship between linearly related Variables

The following formula is Used to Calculate the pearson r Correlation.

$$r = \frac{N\Sigma XY - \Sigma x \Sigma Y}{\sqrt{[N\Sigma x^2 - \Sigma(x^2)][N\Sigma y^2 - \Sigma(y)^2]}}$$

r = pearson r Correlation Co-efficient

N = number of observations

ΣXY = Sum of the products of paired Scores

ΣX = Sum of x Scores

ΣY = Sum of Y Scores

Σx^2 = Sum of Squared x Scores

ΣY^2 = Sum of Squared Y Scores

Kendall rank Correlation: Kendall rank Correlation is a non-parametric test that measures the strength of dependence between two Variables If we Consider two Samples, a and b where each Sample Size is n, we know that the total number of pairings ab is $n(n-1)/2$

$$\tau = \frac{n_c - n_d}{\frac{1}{2} n(n-1)}$$

N_c = number of Concordant

N_d = Number of discordant

Spearman-rank Correlation:

It is a non-parametric test that is Used to measure the degree of association between two Variables