Non Parametric Methods

A large number of statistical methods are available that comprises procedures not requiring the estimation of the population variance or population mean and is not mentioning the hypothesis about the parameters. These testing procedures are known as non parametric test. These methods do not make assumption about the nature of the distribution of the population and hence they are known as distribution free test. Some of the Non parametric methods are

1-sample sign test: Use this test to estimate the median of a population and compare it to a reference value or target value.

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 (like the Cauchy distribution or uniform distribution).

Kruskal-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.

The **Mann-Kendall Trend Test** looks for trends in time-series data.

Mann-Whitney test: Use this test to compare differences 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 test: Use when you want to find a correlation between two sets of data.

Here, we are going to study about the following non parametric methods

1. Mann Whitney U test
2. Kruskal's walli's test
3. Test of significance for differences among several medians

Mann Whitney U test

It is a non-parametric method used to determine whether two independent samples have been drawn from populations with same distribution. This test is also known as U-Test. This test enables us to test the null hypothesis Ho: $\mu_1 = \mu_2$, without assuming whether the population samples have roughly the shape of a normal distribution.

This method helps us to determine whether the two samples have come from identical populations. If it is true that the samples have come from the same population, it is reasonable to assume that the means of ranks assigned to the values of two samples are more or less the same. The alternative hypothesis H_1 would be: That the means of the populations are not equal, i.e., $H_1: \mu_1 \neq \mu_2$. In this case, most of the smaller ranks will go to the values of one sample, while most of the higher ranks will go to the other sample.

Small Sample U-Test

1. Combine the two random samples of size n_1 and n_2 and rank values from smallest to largest. If several values are tied, then assign each the average of the ranks that would otherwise have been assigned. If the two sample sizes are unequal, then suppose n_1 represent smaller-sized sample and n_2 represent the larger-sized sample. The rank sum test statistic U_1 is the sum of the ranks assigned to the n_1 observations in the smaller sample. However, for equal-sized samples, either group may be selected for determining U. The test statistic U_1 plus the sum of the ranks assigned to the n_2 observations in the larger sample $U_1 + U_2 = n (n + 1) / 2$ represents the sum of first consecutive integers.

2. State the null and alternative hypotheses for U-test as follows:

$H_0 : \mu_1 = \mu_2 \leftarrow$ Two populations distribution have equal mean

$H_1 : \mu_1 \neq \mu_2 \leftarrow$ Two populations distribution have different means

The test of the null hypothesis can either be two-tailed or one-tailed.

3. The value of U-statistic is the smallest of the following two U values computed as follows:

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

$$U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2$$

The U-statistic is a measure of the difference between the ranked observations of the two samples. Large or small values of statistic provide evidence of a difference between two populations. If differences between populations are only in location, then large or small values of U-statistic provide evidence of a difference in location (median) of two populations. Both U_1 and U_2 need not be calculated, instead one of U_1 or U_2 can be calculated and other can be formed by using the equation

$U_1 = n_1 n_2 - U_2$.

Large Sample U-Test

For large samples (i.e., when both n_1 and n_2 are greater than 10) the sampling distribution of the U statistic can be approximated by the normal distribution so that z-test statistic is given by $z = \frac{U - \mu_U}{\sigma_U}$ where Mean, $\mu_U = n_1 n_2/2$ and standard deviation, $\sigma_U = \text{sqrt}(n_1 n_2(n_1+n_2+1)/12)$

Decision Rules:

When n_1 and n_2 are both less than or equal to 10, standard table value can be used to obtain the critical value of the test statistic for both one- and two-tailed test at a specified level of significance.

For large sample, at a specified level of significance,

• Reject H_0, if computed value of $z_{cal} \geq$ critical value z_α • Otherwise accept H_0 .

Example: The nicotine contents of two brands of cigarettes, measured in milligrams, was found to be as follows:

Brand A	**2.1**	**4.0**	**6.3**	**5.4**	**4.8**	**3.7**	**6.1**	**3.3**		
Brand B	4.1	0.6	3.1	2.5	4.0	6.2	1.6	2.2	1.9	5.4

Test the hypothesis, at the 0.05 level of significance, that the average nicotine contents of the brands are equal against the alternative that they are unequal.

Solution.

1. Setting up of Hypothesis

Null Hypothesis: $H_0 : \mu_1 = \mu_2$, i.e., the average nicotine contents of the two brands are equal.

Alternative Hypothesis: $H_1 : \mu_1 \neq \mu_2$, i.e., the average nicotine contents of the two brands are not equal.

It is a case of two tailed test.

2. Level of significance: Here $\alpha = 0.05$.

3. Computation of U-statistic:

The observations are arranged in ascending order and ranks from 1 to 18 are assigned

Brand A	**2.1**	**4.0**	**6.3**	**5.4**	**4.8**	**3.7**	**6.1**	**3.3**		
Ranks	4	10.5	18	14.5	13	9	16	8		
Brand B	4.1	0.6	3.1	2.5	4.0	6.2	1.6	2.2	1.9	5.4
Ranks	12	1	7	6	10.5	17	2	5	3	14.5

Now $R_1 = 4 + 10.5 + 18 + 14.5 + 13 + 9 + 16 + 8 = 93$

$R_2 = 12 + 1 + 7 + 6 + 10.5 + 17 + 2 + 5 + 3 + 14.5 = 78$

Also, $n_1 = 8$ and $n_2 = 10$

$$U = n_1 n_2 + \frac{n_1(n_1 + 1)}{2} - R_1 = 23$$

Mean of U, $\mu_U = n_1 n_2/2 = (10 \times 8)/2 = 40$

Variance of U = $\sigma_U^2 = [(n_1 \times n_2)(n_1 + n_2 + 1)] / 12 = 126.67$

$\sigma_u = \sqrt{126.67} = 11.25$. Here $n_2 = 10$, so we can use the statistic $z = \frac{U - \mu_U}{\sigma_U}$

$= (23 - 40) / 11.25 = -1.51.$

The tabled value of Z, at $\alpha = 0.05$ is 1.96. Now, $|Z1| < Z_\alpha$ as $1.51 < 1.06$, so we accept the null hypothesis H_0, and conclude that there is no significant difference in the average nicotine contents of the two brands of cigarettes.

Example: The following are the weight gains (in pounds) of two random samples of young Indians fed on two different diets but otherwise kept under identical conditions:

Diet I: 16.3 10.1 10.7 13.5 14.9 11.8 14.3 10.2

12.0 14.7 23.6 15.1 14.5 18.4 13.2 14.0

Diet II: 21.3 23.8 15.4 19.6 12.0 13.9 18.8 19.2

15.3 20.1 14.8 18.9 20.7 21.1 15.8 16.2

Use U test at 0.01 level of significance to test the null hypothesis that the two population samples are identical against the alternative hypothesis that on the average the second diet produces a greater gain in weight.

Solution.

1. Setting up of Hypothesis:

Null Hypothesis: $H_0 : \mu_1 = \mu_2$

Alternative Hypothesis: : $H_0 : \mu_1 < \mu_2$

2. Level of significance: $\alpha = 0.01$

3. Test Statistic: Ranking the data jointly according to size we find the values first sample occupy the ranks: 21, 1, 3, 8, 15, 4, 11,2, 5.5, 13, 31, 16, 12, 22, 7 and 10 (the fifth and sixth values are both 12, so we assigned each the rank 5.5).

R_1 = The sum of the ranks of the first sample

= 21 +1+3+8+15+4+11+2+5.5+13+31+ 16+ 12+22+7+ 10 = 181.5.

Also $n_1 = 16$ and $n_2 = 16$.

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

= 16 x 16 + 16(16+1) / 2 - 181.5 = 210.5

As the sample sizes n_1 and n_2 are both greater than 8, so the distribution of U is nearly normal with Mean μ_U = 16x 16 /2 = 128.

Variance = $\sigma_U^2 = [(n_1 \times n_2)(n_1 + n_2 + 1)] / 12 = 16x16\ (16+17+1)/12 = 704$

S.D. (U) = $\sigma_U = \sqrt{704} = 26.53$.

$$z = \frac{U - \mu_U}{\sigma_U} = 3.11$$

The test statistic

Critical value. From the normal tables the value of Z, for $\alpha = 0.01$ is 2.33.

4. Decision: Reject the hypothesis if the calculated value of Z is more than the tab value of Z . Here 3.11 > 2.33, so the null hypothesis is rejected and we conclude that the average the second diet produces a greater gain in weight.

KRUSKAL-WALLIS TEST OR H-TEST

The Kruskal-Wallis test is a generalization of Mann-Whitney U-test to the case of k samples. This test is also known as Kruskal-Wallis H-test. It is used to test the null hypothesis H_0, that k independent samples are drawn from the identical population.

The test is an alternative non-parametric test to the F-test for testing the equality of means the one factor analysis of variance when the experimenter wishes to avoid the assumption that the samples were selected from the normal populations. Let n_i, (i = 1, 2, 3,..... k) be number of observations in the i^{th} sample. First we combine all k samples and arrange them to $n = n_1 + n_2 + \ldots, + n_k$ observations in ascending order, substituting the appropriate rank for 1, 2, …, n for each observation.

In the case of ties (identical observations), we follow usual procedure of replacing the observations by the means of the ranks that observations would have if they were distinguishable. The sum of the ranks corresponding the n_i observations in the i^{th} sample is denoted by the random variable R_i. Now let us consider the statistic

$$H = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(n+1)$$

which is approximated very well by a chi-square distribution with k - 1 degrees of freedom when H_0, is true and if each sample consists of at least five observations.

However, if one or more samples have two or more equal observations, the value of H is adjusted as H′ = H/C, where C is the correction factor defined as

$$C = 1 - \frac{1}{n^3 - n} \sum_{j=1}^{r} (o_j{}^3 - o_j)$$

where O_j = the number of equal observations in the j^{th} sample; r = the number of samples which has equal observations. The null and alternative hypotheses are stated as H_0 : The k different populations have identical distribution

H_1 : At least one of the k populations has different distribution

Decision Rule

Reject H_0 when computed value of H is greater than χ^2 (Chi-square) at df = k – 1 and α level of significance . Otherwise accept H_0

Example: Use Kruskal–Wallis test to determine whether there is a significant difference in the following populations. Use α = 0.05 level of significance

Population 1 :	17	19	27	20	35	40	
Population 2 :	28	36	33	22	27		
Population 3 :	37	30	39	42	28	25	31

Solution: Three populations are considered for study, so k = 3 and n = 18. The observations in three populations are combined and ranked. The smallest value is given rank 1, as shown below

Population 1 :	17	19	27	20	35	40
Rank	1	2	6.5	3	13	17

Population 2 :	28	36	33	22	27		
Rank	8.5	14	12	4	6.5		
Population 3 :	37	30	39	42	28	25	31
Rank	15	10	16	18	8.5	5	11

Suppose H_0 : Three populations are identical, i.e., $\mu_1 = \mu_2 = \mu_3$.

$H_1 : \mu_1 \neq \mu_2 \neq \mu_3$

Kruskal–Wallis test statistic is H′ = H/ C ,

$$H = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(n+1)$$

= 0.035[(301.4 + 405 + 996.03)] – 57 = 2.572

$$C = 1 - \frac{1}{n^3 - n} \sum_{j=1}^{r} (o_j^3 - o_j)$$

= 1- (12 /5814) = 1 – 0.002 = 0.998

Thus, H′ = 2.572/ 0.998 = 2.577. Since computed value of H′ (= 2.577) is less than table value of χ^2 (= 5.99) at df = k – 1 = 2 and α = 0.05, the null hypothesis is accepted and it is concluded that three populations are identical.

Example 2: A shoe company wants to know if three groups of workers have different salaries:

Women:	23K	41K	54K	66K	78K
Men:	45K	55K	60K	70K	72K
Minorities:	18K	30K	34K	40K	44K.

Sol:

Step 1: Sort the data for all groups/samples into ascending order in one combined set.

Step 2: Assign ranks to the sorted data points**.** Give tied values the average rank.

Women :	2	6	9	12	15	$R_1 = 44$
Men:	8	10	11	13	14	$R_2 = 56$
Minorities:	1	3	4	5	7	$R_3 = 20$

Step 3: Add up the different ranks for each group/sample**.**

Step 4: Calculate the H statistic

$$H = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{R_i^{\,2}}{n_i} - 3(n+1)$$

where n = sum of sample sizes for all samples,

k = number of samples,

R_i = sum of ranks in the i^{th} sample,

n_i = size of the i^{th} sample.

H = (12/ (15x 16)) ([(44^2/5) + (56^2/5) + (20^2/5)] – (3 x 16)) = 6.72

Step 5: Find the critical chi-square value, with k-1 degrees of freedom. For 3 – 1 degrees of freedom and an alpha level of .05, the critical chi square value is 5.9915

Step 6: Compare the H value from Step 4 to the critical chi-square value from Step 5. If the critical chi-square value is **less than** the H statistic, reject the null hypothesis that the medians are equal. If the chi-square value is **not less** than the H statistic, there is not enough evidence to suggest that the medians are unequal.

In this case, 5.9915 is less than 6.72, so you can reject the null hypothesis.

Test of significance for differences among several medians

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. The alternative hypothesis can be either that the two medians are different (two-tailed test) or that one median is greater than the other (one-tailed test).

The principle of the test is that if two samples have the same median, they should have more or less the same proportion of observations above and below that median. This would be true irrespective of their two distributions. If any scores fall at the value of the combined median they may either be dropped from the analysis, or included with scores less than the median.

The advantage of median test over Mann Whitney is that it only tests for differences in the median irrespective of any differences in the shape of the distribution.

This test can be approximated to chi-square test and the table value can be determined from chi-square table.

Example:

A total of 48 seedlings were planted along four sides of building. The heights after several years of growth are given in the following data. Test the hypothesis that the median height is the same on all four sides of building

North	EAST	South	WEST
7.1	6.9	7.8	6.4
7.2	7	7.9	6.6
7.4	7.1	8.1	6.7
7.6	7.2	8.3	7.1
7.6	7.3	8.3	7.6
7.7	7.3	8.4	7.8
7.7	7.4	8.4	8.2
7.9	7.6	8.4	8.4
8.1	7.8	8.6	8.6
8.4	8.1	8.9	8.7
8.5	8.3	9.2	8.8
8.8	8.5	9.4	8.9

Sol :

Null hypothesis H_0: The medians are equal

Alternative hypothesis H_1: The medians are not equal

$M_1 = 7.7$; $M_2 = 7.35$; $M_3 = 8.4$; $M_4 = 8$

Grand mean $M = (7.7+7.35+8.4+8)/4 = 7.86 = 7.9$ approximately

O Table

	North	East	South	West	Total
No. of values above M	4	3	10	6	23
Below M	8	9	2	6	25
Total	12	12	12	12	48

E Table

	North	East	South	West	Total
No of values above M	5.75	5.75	5.75	5.75	23
No of values below M	6.25	6.25	6.25	6.25	25
Total	12	12	12	12	48

Chi square statistic $\chi^2 = \sum \frac{(O-E)^2}{E} = 9.63$

The table value of chi square is 7.815. Since the calculated value is greater than the table value, we reject the null hypothesis.

Example 2: A major wheat supplier from Texas analyzing the yields of various crop methods. He randomly assigned two different wheat crop methods to a very high number of different acres of farm land and recorded the production rate (yield per acre) for each plot.

Method 1	**1**	**14**	**19**	**12**	**11**	**15**	**20**	**5**	**21**	**15**	**15**	**28**	**3**	**6**
Method 2	16	17	19	10	31	22	26	24	27	32	14	8	12	11

Determine whether there is a significant difference between the two wheat crop methods

Sol:

Null hypothesis H_0: The medians are equal

Alternative hypothesis H_1: The medians are not equal

$M_1 = 15$; $M_2 = 15$

Grand mean $M = (15 + 15)/2 = 15$

	Method 1	Method 2	Total
Above M	4	9	13
Below M	10	5	15
Total	14	14	28

Expected Frequencies :

Expected value for each cell= row total * column total/grand total

E Table

	Method 1	Method 2	Total
Above M	6.5	6.5	13
Below M	7.5	7.5	15
Total	14	14	28

$$\chi^2 = \sum \frac{(O-E)^2}{E} = 3.56$$

Compute the degrees of freedom = (2-1)*(2-1) =1

Critical value of χ^2 (0.05, 1) = 3.841

Calculated χ^2 value is less than the critical value of χ^2 for a 0.05 significance level, hence we have no convincing evidence to reject the null hypothesis

So, the data are consistent with the null hypothesis that there is no difference between the two wheat crop methods.