Rank-Sum Test

A non parametric method is one that satisfies at least one of the following criteria

(i)The method deals with enumerative data(data that are frequently counts)

(ii) the method does not deal with specific population parameters such as μ and σ.

(iii)the method does not nrequire assumptions about specific population distributions.(in particular, the assumption of normality)

The assumptions associated with non-parametric tests are

(i)sample observations are independent

(ii)the variable under study is continuous

(iii)Lower order moments exist.

Advantages of Non-Parametric Methods

(i)They do not require us to make the assumption that a population is distributed in the shape of a normal curve or another specific shape

(ii)Generally they are easier to do and to understand.

(iii)Sometimes even formal ordering or ranking is not required.

Disadvantages of Non-Parametric Methods

(i) The ignore a certain amount of information
(ii) They are often not as efficient or sharp as parametric tests
(iii) The non-parametric tests cannot be used to estimate parameters in the population (or) the confidence intervals for such parameters
(iv) It is not possible to solve certain statistical problems by using non-parametric tests. A good example is the type of problem dealt in the analysis of variance

Uses of Non-Parametric Methods

There are four important situations in which the use of a distribution free or non-parametric technique is indicated

(i)when quick or preliminary data analysis is needed.

(ii) when the assumptions of a competing distribution-tied or parametric procedure are not satisfied and the consequences of this are either unknown or known to be serious.

(iii) when data are only roughly scaled, for example when only comparative rather than absolute magnitudes are available

(iv)when the basic questions of interest if distribution- free or non parametric in nature. For example are these two samples drawn from populations with identical distributions?

Rank-Sum Test

sign test is a non-parametric statistical test for identifying differences between two populations based on the analysis of nominal data.

But the rank sum test is a non- parametric test for identifying differences between two or more populations based on the analysis of two or more independent samples one from each population are used. We shall concentrate only on the following two tests.

1.Mann-Whitney U-test

2. Kruskal-Wallis test or H-test

Mann-Whitney U-test and Kruskal-Wallis test are called rank-sum tests because the test depends on the ranks of the sample observations. Mann-Whitney U-test is used when there are only two populations whereas Kruskal-Wallis test is employed when more than two populations are involved.

Mann-Whitney U-test

Use of Mann-Whitney U-test will enable us to determine whether the two populations are identical. Let $\{x_1, x_2, x_{3,} \dots \dots x_m\}$ and $\{y_1, y_2, y_{3,} \dots \dots y_n\}$ be two independent random samples from two populations. Here we set up the null hypothesis

Two populations are identical

$$H_0: \mu_1 = \mu_2$$

Two populations are not identical

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

Working Rule

Step:1

Combine all the given samples (from smallest to largest), and assign ranks to all these values.

Step:2

Assign the average of the rank if the sample values are same.

Step:3

Find the sum of the ranks for each of the sample. Let us denote these sum by $R_1$and R_2. Also n_1 and n_2 are their respective sample sizes.

Step:4

Calculate 'U"- Statistic:

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

OR

$$U = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2 \text{ (For Sample-2)}$$

The mean of U are mean $E(U) = \frac{n_1 n_2}{2}$.

Variance of 'U' are $V(U) = \frac{n_1 n_2 (n_1+n_2+1)}{12}$

The standard normal variant of 'U' are $Z = \frac{U - E(U)}{\sqrt{V(U)}}$

Step:5

If $|Z| \leq Z_\alpha$ are accept H_0 and reject H_0 when $|Z| > Z_\alpha$, where Z_α is the table value of Z for the level of significance α.

Problems

1. The nicotine con tents 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

Solution:

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

The average nicotine contents of the two brands are equal

Alternative Hypothesis: $H_1: \mu_1 \neq \mu_2$

The average nicotine contents of the two brands are not equal

Level of significance: $\alpha = 0.05$

Brand A	Rank R_1	Brand B	Rank R_2
2.1	4	4.1	12
4.0	10.5	0.6	1
6.3	18	3.1	7
5.4	14.5	2.5	6
4.8	13	4.0	10.5
3.7	9	6.2	17
6.1	16	1.6	2
3.3	18	2.2	5
		1.9	3
		5.4	14.5
	$\sum R_1 = 93$		$\sum R_2 = 78$

$$n_1 = 8, n_2 = 10$$

Calculate U-Statistics

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

$$U = 8 * 10 + \frac{8(8 + 1)}{2} - 93$$

$$= 80 + \frac{72}{2} - 93 = 80 + 36 - 93 = 23$$

$$U = 23$$

Mean of U

$$\mu = E(U) = \frac{n_1 n_2}{2}$$

$$= \frac{8 * 10}{2} = 40$$

$$\mu = E(U) = 40$$

Variance of ‘U’

$$V(U) = \frac{n_1 n_2 (n_1 + n_2 + 1)}{12}$$

$$= \frac{8 * 10(8 + 10 + 1)}{12} = \frac{80 * 19}{12} = 126.67$$

$$V(U) = 126.67$$

$$U = 23, \quad \mu = E(U) = 40, \quad V(U) = 126.67$$

The standard normal variant of ‘U’

$$Z = \frac{U - E(U)}{\sqrt{V(U)}}$$

$$= \frac{23 - 40}{\sqrt{126.67}} = \frac{-17}{\sqrt{126.67}}$$

$$= -1.51$$

$$|Z| = 1.51$$

At $\alpha = 0.05, \;\; Z_\alpha = 1.96$

Conclusion

$|Z| < Z_\alpha$

We accept H_0.

2. From a maths class of 12 equally capable students using a programmed material, five are selected at random and given additional instructions by the teacher. The results on the final exams is as follows.

Additional Instruction	87	69	78	91	80		
No Additional Instruction	75	88	64	82	93	79	67

Use the rank sum test at 5% level of significance to determine if the additional instruction affects the average grade.

Solution:

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

i.e) there is no significant difference in the average grade.

Alternative Hypothesis: $H_1: \mu_1 \neq \mu_2$

i.e) there is a significant difference in the average grade

Level of significance: $\alpha = 0.05$

Additional Instruction	Rank R_1	No Additional Instruction	Rank R_2
87	9	75	4
69	3	88	10
78	5	64	1
91	11	82	8
80	7	93	12
		79	6
		67	2
	$\sum R_1 = 35$		$\sum R_2 = 43$

$$n_1 = 5,\ n_2 = 7,\ \sum R_1 = 35,\ \sum R_2 = 43$$

Calculate U-Statistics

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

$$U = 5 * 7 + \frac{5(5+1)}{2} - 35$$

$$= 35 + \frac{30}{2} - 35 = 35 + 15 - 35 = 15$$

$$U = 15$$

Mean of U

$$\mu = E(U) = \frac{n_1 n_2}{2}$$

$$= \frac{5 * 7}{2} = 17.5$$

$$\mu = E(U) = 17.5$$

Variance of 'U'

$$V(U) = \frac{n_1 n_2 (n_1 + n_2 + 1)}{12}$$

$$= \frac{5 * 7(5+7+1)}{12} = \frac{35 * 13}{12} = 37.91$$

$$V(U) = 37.91$$

$$U = 15, \quad \mu = E(U) = 17.5, \quad V(U) = 37.91$$

The standard normal variant of 'U'

$$Z = \frac{U - E(U)}{\sqrt{V(U)}}$$

$$= \frac{15 - 17.5}{\sqrt{37.91}} = \frac{-2.5}{\sqrt{37.91}}$$

$$= -0.406$$

$$|Z| = 0.406$$

At $\alpha = 0.05, \;\; Z_\alpha = 1.96$

Conclusion

$|Z| < Z_\alpha$

We accept H_0.

3. Two methods of instruction to apprentices is to be evaluated. A director assigns 15 randomly selected trainees to each of the two methods. Due to drop outs, 14 complete in batch 1 and 12 complete in batch 2. An achievement test was given to these successful candidates. Their scores are as follows.

   Method I: 70, 90,82,64,86,77,84,79,82,89,73,81,83,66
   Method II: 86,78,90,82,65,87,80,88,95,85,76,94

   Test whether the two methods have significant difference in effectiveness. Use Mann-Whitney test at 5% significant level.

Solution:

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

i.e) there is no significant difference in effectiveness between the two methods.

Alternative Hypothesis: $H_1: \mu_1 \neq \mu_2$

i.e) there is a significant difference in effectiveness between the two methods

Level of significance: $\alpha = 0.05$

Model I	Rank R_1	Model II	Rank R_2
70	4	86	18.5
90	23.5	78	8
82	13	90	23.5
64	1	82	13
86	18.5	65	2
77	7	87	20
84	16	80	10
79	9	88	21

82	13	95	26
89	22	85	17
73	5	76	6
81	11	94	25
83	15		
66	3		
	$\sum R_1 = 161$		$\sum R_2 = 190$

$n_1 = 14,$ $n_2 = 12,$

$\sum R_1 = 161,\ \sum R_2 = 190$

Calculate U-Statistics

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

$$U = 14 * 12 + \frac{14(14+1)}{2} - 161$$

$$= 168 + 105 - 161$$

$$U = 112$$

Mean of U

$$\mu = E(U) = \frac{n_1 n_2}{2}$$

$$= \frac{14 * 12}{2} = 84$$

$$\mu = E(U) = 84$$

Variance of ‘U’

$$V(U) = \frac{14 * 12(14+12+1)}{12}$$

$$= 14 * 27 = 378$$

$$V(U) = 378$$

$$U = 112, \qquad \mu = E(U) = 84, \quad V(U) = 378$$

The standard normal variant of ‘U’

$Z = \frac{U - E(U)}{\sqrt{V(U)}}$

$$= \frac{112 - 84}{\sqrt{378}} = \frac{28}{\sqrt{378}}$$

$$= 1.4402$$

$$|Z| = 1.4402$$

At $\alpha = 0.05, \ Z_\alpha = 1.96$

Conclusion

$|Z| < Z_\alpha$

We accept H_0.

Home work

1. The following are the weight gains(in pounds) of two random samples of young Indian 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	1.2
	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.

2. The following random samples are measurements of the heat producing capacity (in millions of calories per ton) of specimens of coal from two mines

Mine I :	31	25	38	33	42	40	44	26	43	35
Mine II :	44	30	34	47	35	32	35	47	48	34

Test the hypothesis of no difference between the Mine I and Mine II.

Using the Mann-Whitney 'U' test for the above sample data. Use $\alpha = 0.10$

3. The following are the number of mistakes counted on pages randomly selected from reports typed by a company's two secretaries.

Male Secretary	:	15	10	5	6	8	10	12
Female Secretary	:	12	8	7	9	10	5	4

Use 'U' test at 2% level of significance to test the null hypothesis that the 2 secretaries average equal mistakes per page.

Kruskal-Wallis test or H-test

The Mann-Whitney U test can be used to test whether two populations are identical. It has been extended to the case of 3 or more populations by Kruskal-Wallis .

The hypothesis K-W test with $k > 3$ populations can be written as

Null Hypothesis: $H_0: \mu_1 = \mu_2 = \mu_3$

i.e) all the populations are identical

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

i.e) all the populations are not identical

The K-W test which is based on the sum of the ranks for each of the samples can be computed as follows.

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

Where $n_i =$ the number of items in sample i

$k =$ Number of populations(or samples)

$n = \sum n_i = n_1 + n_2 + \cdots . + n_k.$

i.e) the total number of observations in all samples

R_i= sum of the ranks of all items in sample i

To compute the W statistic , we must first rank all the given sample items. Kruskal-Wallis show that under the null hypothesis in which the populations are identical, the sampling distribution of ‘W’ can be approximately by a χ^2 distribution with $(k-1)$ df.

The approximation is generally acceptable if each of the sample sizes is greater than or equal to 5.

If H falls in the critical region $H \leq {\chi_\alpha}^2$ with $(k-1)$ degrees of freedom, we accept our null hypothesis at α level of significance, otherwise we reject H_0.

PROBLEMS:

1. Use Kruskal-Wallis to test for differences in mean among the 3 samples. If $\alpha = 0.01$, what are your conclusions.

Sample I : 95 97 99 98 99 99 99 94 95 98

Sample II : 104 102 102 105 99 102 111 103 100 103

Sample III : 119 130 132 136 141 172 145 150 144 135

Solution

3 samples are given $\Longrightarrow k = 3$

Null Hypothesis: $H_0: \mu_1 = \mu_2 = \mu_3$

i.e) there is no difference in mean among the 3 samples

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

i.e) there is a difference in mean among the 3 samples

Level of significance: $\alpha = 0.01$

Sample I	Rank R_1	Sample II	Rank R_2	Sample III	Rank R_3
95	2.5	104	18	119	21
97	4	102	14	130	22
99	9	102	14	132	23
98	5.5	105	19	136	25
99	9	99	9	141	26
99	9	102	14	172	30
99	9	111	20	145	28
94	1	103	16.5	150	29
95	2.5	100	12	144	27
98	5.5	103	16.5	135	24

	$\sum R_1 = 57$		$\sum R_2 = 153$		$\sum R_3 = 255$

$$n = n_1 + n_2 + n_3 = 10 + 10 + 10 = 30$$

The test statistics is,

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

$$= \frac{12}{n(n+1)}\left[\frac{{R_1}^2}{n_1} + \frac{{R_2}^2}{n_2} + \frac{{R_3}^2}{n_3}\right] - 3(n+1)$$

$$= \frac{12}{30 * 31}\left[\frac{57^2}{10} + \frac{153^2}{10} + \frac{255^2}{10}\right] - 3(30+1) = 25.30$$

$$W \; or \; H = 25.30$$

The χ^2 value at 1% level with $(k-1) = 3 - 1 = 2$ degrees of freedom is ${\chi_\alpha}^2 = 9.21$

Conclusion:

$W \; or \; H > {\chi_\alpha}^2$

We reject H_0

2. A research company has designed three different systems to clean up oil spills. The following table contains the results, measured by how much surface area(in square meters) is cleaned in one hour. The data were found by testing each method in several trials. Are three systems equally effective? Use the 5% level of significance.

System	A :	55	60	63	56	59	55
System	B :	57	53	64	49	62	
System	C :	66	52	61	57		

Solution

3 systems are given $\Longrightarrow k = 3$

Null Hypothesis: $H_0 : \mu_1 = \mu_2 = \mu_3$

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

Level of significance: $\alpha = 0.05$

System A	Rank R_1	System B	Rank R_2	System C	Rank R_3
55	4.5	57	7.5	66	15
60	10	53	3	52	2
63	13	64	14	61	11
56	6	49	1	57	7.5
59	9	62	12		
55	4.5				
	$\sum R_1 = 47$		$\sum R_2 = 37.5$		$\sum R_3 = 35.5$

$$n_1 = 6, \;\; n_2 = 5, n_3 = 4$$

$$n = n_1 + n_2 + n_3 = 6 + 5 + 4 = 15$$

The test statistics is,

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

$$= \frac{12}{n(n+1)} \left[\frac{{R_1}^2}{n_1} + \frac{{R_2}^2}{n_2} + \frac{{R_3}^2}{n_3} \right] - 3(n+1)$$

$$= \frac{12}{15 * 16} \left[\frac{47^2}{6} + \frac{37.5^2}{5} + \frac{35.5^2}{4} \right] - 3(15+1) = 0.224$$

$$W \; or \; H = 0.224$$

The χ^2 value at 5% level with $(k-1) = 3 - 1 = 2$ degrees of freedom is ${\chi_\alpha}^2 = 5.991$

Conclusion:

$W \; or \; H < {\chi_\alpha}^2$

We accept H_0.

3. The Molisa's shop has 3 mall locations. She keeps a daily record for each location of the number of customers who actually make a purchase. A sample of these data follows. Using Kruskal-Wallis test can you say that 5% level of significance that her stores have the same number of customers who buy.

Eastowin : 99 64 101 85 79 88 97 95 90 100
Craborchard : 83 102 125 61 91 96 94 89 93 75
Fair Forest : 889 98 56 105 87 90 87 101 76 89

Solution

3 systems are given $\Longrightarrow k = 3$

Null Hypothesis: $H_0: \mu_1 = \mu_2 = \mu_3$

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

Level of significance: $\alpha = 5\%$

Eastowin	Rank R_1	Craborchard	Rank R_2	Fair Forest	Rank R_3
99	24	83	7	89	13
64	3	102	28	98	23
101	26.5	125	30	56	1
85	8	61	2	105	29
79	6	91	17	87	9.5
88	11	96	21	90	15.5
97	22	94	19	87	9.5
95	20	89	13	101	26.5
90	15.5	93	18	76	5
100	25	75	4	89	13
	$\sum R_1 = 161$		$\sum R_2 = 159$		$\sum R_3 = 145$

$$n_1 = 10, \;\; n_2 = 10, n_3 = 10$$

$$n = n_1 + n_2 + n_3 = 10 + 10 + 10 = 30$$

The test statistics is,

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

$$= \frac{12}{n(n+1)}\left[\frac{{R_1}^2}{n_1} + \frac{{R_2}^2}{n_2} + \frac{{R_3}^2}{n_3}\right] - 3(n+1)$$

$$= \frac{12}{30*31}\left[\frac{161^2}{10} + \frac{159^2}{10} + \frac{145^2}{10}\right] - 3(30+1) = 0.196$$

$$W \text{ or } H = 0.196$$

The χ^2 value at 5% level with $(k-1) = 3-1 = 2$ degrees of freedom

is ${\chi_\alpha}^2 = 5.991$

Conclusion:

$W \text{ or } H < {\chi_\alpha}^2$

We accept H_0.

Home Work

1. A company's trainees are randomly assigned to groups which are taught a certain industrial inspection procedure by 3- different methods. At the end of the instruction period they are tested for inspection performance quality. The following are their scores.

Method A :	80	83	79	85	90	68	
Method B	82	84	60	72	86	67	91
Method C	93	65	77	78	88		

Use H test to determine at the 0.05 level of significance whether the 3 methods are equally effective.

2. An information systems company investigated the computer literacy of managers. As a part of their study, the company designed a questionnaire. To check the design of the questionnaire (ie. Its validity), 19 managers were randomly selected and asked to complete the questionnaire. The managers were classified as A, B and C based on their knowledge and experience. The

scores appear in the table below. Is there sufficient evidence to conclude that the mean scores differs for the 3- groups of managers? Use $\alpha = 0.05$

	Level	
A	**B**	**C**
82	128	156
114	90	128
90	130	151
80	110	140
88	133	
93	130	
80	104	
105		

3. A quality control engineer in an electronics plant has sampled the output of three assembly lines and recorded the number of defects observed. The samples involve the entire output of the three lines for 10 randomly selected hours from a given week. Do the data provide sufficient evidence to indicate that atleast one of the line tends to produce more defects than the others. Test at the 5% level of significance using suitable non parametric test.

Line 1 :	6	38	3	17	11	30	15	16	25	5
Line 2 :	34	28	42	13	40	31	9	32	39	27
Line 3 :	13	35	19	4	29	0	7	33	18	24