Non Parametric Statistics

Bijay Lal Pradhan, Ph.D.

Frequently Used Nonparametric Tests

Wilcoxon Rank Sum Test

Wilcoxon Signed Rank Test

Mann Whitney U-Test

Kruskal Wallis H-Test

Wilcoxon test

If the data for analysis are measure at interval or ratio scale and donot want to lose information, then appropriate test is Wilcoxon, which makes use of the magnitude of the differences between measurement and a hypothesized location parameter rather than just the signs of difference.

Assumptions

The sample is random

The variable is continuous

Population is symmetrical about its mean

Wilcox test

H0: $\eta = \eta_0$ **(set null and alt hypothesis)**

Ha: $\eta \neq \eta_0$

α = 0.05 **(set level of significance)**

Test Statistic:

Subtract hypothesized value di=xi-η_0 and if di=0, eliminate di from the calculation, Rank the usable di from the smallest to largest without regard to the sign of di, after that find the sum of the positive (T+) and negative (T-) ranks.

Test statistics=Min(T+,T-) for two tail

=T+ (for left test) T- (for right tail test)

P-Value:

(Wilcox Table: *n*, *T*)

P-value=2*.T (For two tail)

or

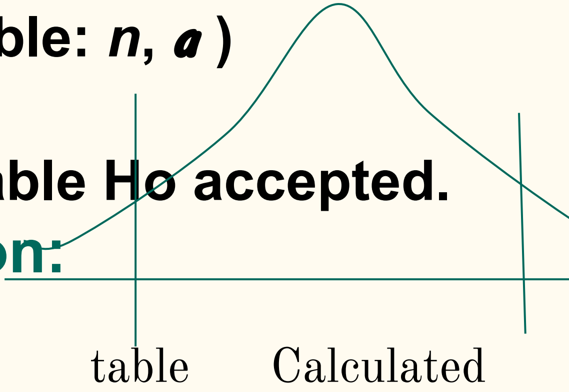
Table Value:

(Wilcox Table: *n*, α)

Decision:

If T cal>t table Ho accepted.

Conclusion:



Wilcoxon Test Example

Suppose you are a botanist studying the heights of a particular species of plant. You have collected a sample of 10 plants and measured their heights.

Measurement of heights in cm= [16, 14, 13, 15, 17, 18, 19, 12, 23, 24]

Test whether the median height of this species of plant is different from 15 cm.

If the data doesn't follow normality.

Wilcox test

H_0: $\eta = 15$ (set null and alt hypothesis)
H_a: $\eta \neq 15$
$\alpha = 0.05$ (set level of significance)

Test Statistic:

$[16, 14, 13, 15, 17, 18, 19, 12, 23, 24]$

Test statistics=Min(T+,T-) =

Table Value:

(Wilcox Table: *n*, *a*)

Decision:

Conclusion:

Wilcox test

H_0: $\eta = 15$ (set null and alt hypothesis)

H_a: $\eta \neq 15$

$\alpha = 0.05$ (set level of significance)

Test Statistic:

Subtract hypothesized value di=xi-15

[16, 14, 13, 15, 17, 18, 19, 12, 23, 24]

[1, -1, -2, 0, 2, 3, 4, -3, 8, 9]

Test statistics=Min(T+,T-) =

Table Value:

(Wilcox Table: *n*, *α*)

Decision:

Conclusion:

Wilcox test

H_0: $\eta = 15$ (set null and alt hypothesis)

H_a: $\eta \neq 15$

$\alpha = 0.05$ (set level of significance)

Test Statistic:

Subtract hypothesized value di=xi-15

[16, 14, 13, 15, 17, 18, 19, 12, 23, 24]

[1, -1, -2, 0, 2, 3, 4, -3, 8, 9]

Rank [1.5, -1.5, -3.5, 3.5, 5.5, 7, -5.5, 8, 9]

T- = 10.5 T+=34.5

Test statistics=Minimum(T+,T-) = 10.5

Table Value:

(Wilcox Table: *n*, *a*)

Decision:

Conclusion:

Wilcox test

H_0: η = 15 (set null and alt hypothesis)

H_a: η ≠ 15

α = 0.05 (set level of significance)

Test Statistic:

Subtract hypothesized value di=xi-15

[16, 14, 13, 15, 17, 18, 19, 12, 23, 24]

[1, -1, -2, 0, 2, 3, 4, -3, 8, 9]

Rank [1.5, -1.5, -3.5, 3.5, 5.5, 7, -5.5, 8, 9]

T- = 10.5 T+=34.5

Test statistics=Min(T+,T-) = 10.5

Table Value:

(Wilcox Table: *9*, *0.05*)=6

Decision:

Conclusion:

Wilcox test

H_0: η = 15 (set null and alt hypothesis)

H_a: η ≠ 15

α = 0.05 (set level of significance)

Test Statistic:

Subtract hypothesized value di=xi-15

[16, 14, 13, 15, 17, 18, 19, 12, 23, 24]

[1, -1, -2, 0, 2, 3, 4, -3, 8, 9]

Rank [1.5, -1.5, -3.5, 3.5, 5.5, 7, -5.5, 8, 9]

T- = 10.5 T+=34.5

Test statistics=Min(T+,T-) = 10.5

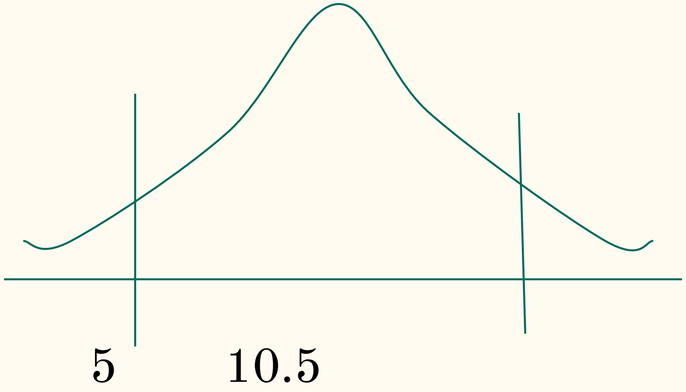
Table Value:

(Wilcox Table: *9*, *0.05*)=5

Decision:

H1 cannot be accepted

Conclusion:



Wilcox test

H_0: η = 15 (set null and alt hypothesis)
H_a: η ≠ 15
α = 0.05 (set level of significance)
Test Statistic:

Subtract hypothesized value di=xi-15

[16, 14, 13, 15, 17, 18, 19, 12, 23, 24]

[1, -1, -2, 0, 2, 3, 4, -3, 8, 9]

Rank [1.5, -1.5, -3.5, 3.5, 5.5, 7, -5.5, 8, 9]

T- = 10.5 T+=34.5

Test statistics=Min(T+,T-) = 10.5

P-Value:
Table Value:

(Wilcox Table: *9*, *0.05*)=6

Decision:

H1 cannot be accepted

Conclusion:

The median value is not significantly different than 15.

Wilcoxon Matched Pair signed Rank test

Suppose you have a group of 10 patients who undergo a new treatment. You measure their blood pressure before and after the treatment to determine if there is a significant change. Here are the before and after treatment blood pressure readings for each patient:

Patient	Before	After
1	120	115
2	130	132
3	115	110
4	140	138
5	135	130
6	125	123
7	150	145
8	155	150
9	160	158
10	145	140

Wilcoxon Matched Pair signed Rank test

Suppose you have a group of 10 patients who undergo a new treatment. You measure their blood pressure before and after the treatment to determine if there is a significant change. Here are the before and after treatment blood pressure readings for each patient:

Patient	Before	After	Di	AD	Rank	Signed Rank
1	120	115	-5	5	7.5	-7.5
2	130	132	2	2	2.5	2.5
3	115	110	-5	5	7.5	-7.5
4	140	138	-2	2	2.5	-2.5
5	135	130	-5	5	7.5	-7.5
6	125	123	-2	2	2.5	-2.5
7	150	145	-5	5	7.5	-7.5
8	155	150	-5	5	7.5	-7.5
9	160	158	-2	2	2.5	-2.5
10	145	140	-5	5	7.5	-7.5

$T+ = 2.5$ $T- = 52.5$ $(1+2+3+4)/4=2.5$ $(5+6+7+8+9+10)/6=7.5$

Wilcoxon Matched Pair signed Rank test

H_0: $\eta_1 = \eta_2$

H_a: $\eta_1 < \eta_2$

$\alpha = 0.05$

Test Statistic:

T+ =2.5 T- =52.5

Test statistics=Min(T+,T-) = 2.5

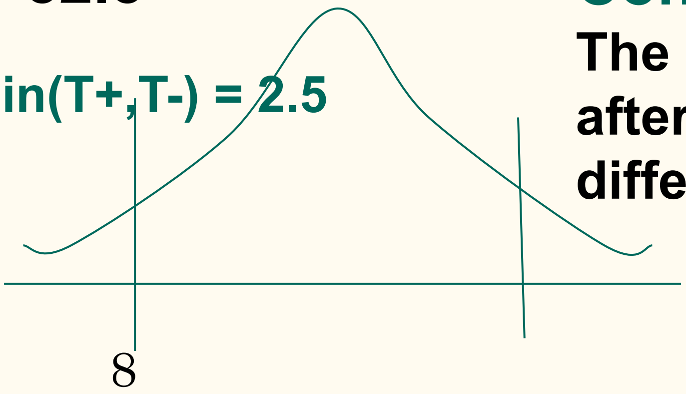


Table Value:

(Wilcox Table: *10*, *0.05*)=8

Decision:

H1 is accepted

Conclusion:

The median value before and after treatment is significantly different.

Mann Whitney U test

Sometimes it is also known as Mann-Whitney-Wilcoxon test (because it is based on the rank sum)

Assumptions

The two sample of size n1 and n2 respectively available for analysis have been independently and randomly drawn from their respective population.

The measurement scale is at least ordinal

The variable of interest is continuous.

Mann Whitney U test

- **Hypothesis**
  H_0: $Md_1 = Md_2$

  H_1: $Md_1 \neq Md_2$ (two tailed) or H_1: $Md_1 > Md_2$ (one tailed right) or H_1: $Md_1 < Md_2$ (one tailed left)

  Combine n_1 and n_2 such that $n_1+n_2= n$ and rank these n observations in ascending order .If two or more observations are equal then assign average rank and is called tied. Sum the ranks of sample of sizes n_1 and n_2 separately to get R_1 and R_2. If two sample sizes are unequal then smaller one is n_1. Obtain U_1 and U_2 as follows

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1 \text{ and } U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2.$$

- **Test Statistics:** $U_0 = \min\{U_1, U_2\}$
- **Level of significance:** Generally fix $\alpha = 0.05$ unless we are given.
- **Critical value:** At α level of signi. critical value from Mann Whitney table as $p = \text{Prob}(U \leq U_0)$.
- **Decision:** Accept H_0 if $p > \alpha$ for one tailed test and $2p > \alpha$ for two tailed test, reject otherwise.
- **Alternately:** At α level of significance, critical value from Mann Whitney table as $U_{tabulated} = U_{a(n1,n2)}$ for two tail and $U_{a/2(n1,\, n2)}$ for one tail test. Accept H_0 if $U_0 > U_{tabulated}$ reject otherwise.

Example

- The heart beating rate of 5 vegetarians and 5 non vegetarians are recorded below:

Vegetarians	56	67	82	60	75
Non vegetarians	53	42	75	58	65

- Is the mean heart beating rate of non vegetarians significantly high. Use Mann Whitney U test.
- **Solution:**

Vegetarians	Ranks	Non vegetarians	Ranks
56	3	53	2
67	7	42	1
82	10	75	8.5
60	5	58	4
75	8.5	65	6
	$R_1 = 33.5$		$R_2 = 21.5$

H_0: There is no significant difference between heart beating rate of vegetarian and non vegetarian ($Md_1 = Md_2$)
H_1: Heart beating rate of non vegetarian is significantly high than vegetarian ($Md_1 < Md_2$)

- Here, Sample size of vegetarian (n_1) = 5; Sample size of Non vegetarian (n_2) = 5
  Sum of ranks of vegetarian (R_1) = 33.5 Sum of ranks of non vegetarian (R_2) = 21.5
- Calculations:

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1 = 5 \times 5 + \frac{5 \times 6}{2} - 33.5 = 6.5$$

$$U_2 = n_1 n_2 - U_1 = 5 \times 5 - 6.5 = 18.5$$

$$U_0 = \min\{U_1, U_2\} = 6.5$$

- **Critical value:** Here the p-value is more than 0.111. (From table); cannot accept H1 at α=0.05

Krushkal Wallis H Test

It is also called Kruskal Wallis one way ANOVA test

It is test used to test the significant difference of median among three or more independent populations.

Let us consider k independent samples of size n_i such that $\Sigma n_i = n$ drawn from continuous population with unknown medians Md_1, Md_2 ....Md_k respectively.

SN	Samples				
1	x_{11}	x_{12}	x_{13}	x_{1j}	x_1n_1
2	x_{21}	x_{22}	x_{23}	x_{2j}	x_2n_2
3	x_{31}	x_{32}	x_{33}	x_{3j}	x_3n_3
.					
i	x_{i1}	x_{i2}	x_{i3}	x_{ij}	x_in_i
.					
.					
k	x_{k1}	x_{k2}	x_{k3}	x_{kj}	x_kn_k

Problem to test

H_0: $Md_1 = Md_2 = Md_3 =$ .... Md_k

H_1: At least one Md_i is different i = 1, 2, 3, ...k.

Krushkal Wallis H Test

Problem to test

H_0: $Md_1 = Md_2 = Md_3 = \ldots\ldots\ldots\ldots\ldots Md_k$

H_1: At least one Md_i is different $i = 1, 2, 3, \ldots\ldots\ldots k$.

Combine n_1, n_2, n_3 .... and n_k such that $n_1 + n_2 + n_3 + \ldots\ldots\ldots + nk = n$ and rank these n observations in ascending order .If two or more observations are equal then assign average rank and is called tied. Sum the ranks of sample of sizes n_1, n_2, n_3, ... and n_k separately to get R_1, R_2, R_3,......... R_k.

Test statistic
$$H = \frac{12}{n(n+1)} \sum \frac{R_i^2}{n_i} - 3(n+1)$$

If tied occurs then corrected test statistic is,
$$H = \frac{\frac{12}{n(n+1)} \sum \frac{R_i^2}{n_i} - 3(n+1)}{1 - \sum \frac{(t_i^3 - t_i)}{n^3 - n}}$$

t_i = number of times i^{th} rank is repeated.

Decision process

Level of significance

Let α be the level of significance .Generally fix $\alpha = 0.05$ unless we are given.

Critical value

For ni ≤ 5 and $k = 3$, critical value p is obtained from Kruskal Wallis table.

For $n_i > 5$ and $k > 3$, critical value is$\chi^2_{\alpha\ (k\ -1)}$.

Decision

Accept H_0 at α level of significance if $p > \alpha$, reject otherwise for $n_i \leq 5$ and $k=3$.

Reject H0 at α level of significance if $H > \chi^2\alpha\ (k-1)$, accept otherwise for $n_i > 5$ and $k > 3$.

Example

A bacteriologist was interested to study the number of plankton organism inhabiting the lake water. He made hauls of water from three lakes each and the following results were obtained.

Lake	Number of plankton organism				
Phewa	12	19	16		
Rara	4	8	3	2	3
Taudaha	14	12	20	12	

Do the data provide substantial evidence to conclude significant variation between lake water? Use Kruskal Wallis test at 0.05 level of significance.

Solution **Problem to test**

H_0: There is no significant variation between lake water.($Md_1 = Md_2 = Md_3$)

H_1: There is at least one significant variation between lake water. (At least one Mdi is different)

Lake	Number of plankton organism					R_i	R_i^2/n_i
Phewa	12	19	16				
Rank	7	11	10			28	261.33
Rara	4	8	3	2	3		
Rank	4	5	2.5	1	2.5	15	45
Taudaha	14	12	20	12			
Rank	9	7	12	7		35	306.25
							612.58

$$H = \frac{\frac{12}{n(n+1)} \sum \frac{R_j^2}{n_i} - 3(n+1)}{1 - \sum \frac{(t_i^3 - t_i)}{n^3 - n}} = \frac{\frac{12}{12\,(12+1)} \times 612.5 - 3(12+1)}{1 - \frac{30}{12^3 - 12}} = \frac{8.12}{0.98} = 8.285$$

Critical value: From Kruskal Wallis table critical value is $p = 0.01$

Decision: $P = 0.01 < \alpha = 0.05$, reject H_0 at 0.05 level of significance.

Conclusion: There is at least one significant variation between lake water.

Example

The following are the numbers of misprints counted on pages selected at random from three editions of a book

Edition I	4	10	2	6	4	12
Edition II	8	5	13	8	8	10
Edition III	7	9	11	2	14	7

Use Kruskal Wallis H test at the 0.05 level of significance to test the null hypothesis that the samples come from identical populations.

Solution

Problem to test

H_0: Populations are identical. ($Md_1 = Md_2 = Md_3$)

H_1: Populations are not identical.(At least one Md_i is different, i = 1 , 2 , 3)

Date	No. of misprints						R_i	R_i^2/n_i
April 11	4	10	2	6	4	12		
Rank	3.5	13.5	1.5	6	3.5	16	44	322.66
April 18	8	5	13	8	8	10		
Rank	10	5	17	10	10	13.5	65.5	715.041
April 25	7	9	11	2	14	7		
Rank	7.5	12	15	1.5	18	7.5	61.5	630.375
Total								1668.076

Sample size of April 11 (n_1) = 6

Sample size of April 18 (n_2) = 6

Sample size of April 25 (n_3) = 6

Total sample size $(n) = n_1 + n_2 + n_3 = 6 + 6 + 6 = 18$

No of times rank 1.5 is repeated (t_1) = 2,

No of times rank 3.5 is repeated (t_2) = 2,

No of times rank 7.5 is repeated (t_3) = 2,

No of times rank 10 is repeated (t_4) = 3,

No of times rank 13.5 is repeated (t_5) = 2

$$\Sigma(t_i^3 - t_i) = (2^3 - 2) + (2^3 - 2) + (2^3 - 2) + (3^3 - 3) + (2^3 - 2) = 48$$

$$H = \frac{\frac{12}{n(n+1)} \Sigma \frac{R_j^2}{n_i} - 3(n+1)}{1 - \Sigma \frac{(t_i^3 - t_i)}{n^3 - n}} = \frac{\frac{12}{18\,(18+1)} \times 1668.076 - 3(18+1)}{1 - \frac{48}{18^3 - 18}} = \frac{58.528 - 57}{0.9917} = 1.5407$$

Test

$$H = \frac{\frac{12}{n(n+1)} \sum \frac{R_j^2}{n_i} - 3(n+1)}{1 - \sum \frac{(t_i^3 - t_i)}{n^3 - n}} = \frac{\frac{12}{18\,(18+1)} \times 1668.076 - 3(18+1)}{1 - \frac{48}{18^3 - 18}} = \frac{58.528 - 57}{0.9917} = 1.5407$$

Critical value

Critical value at 0.05 level of significance for 2 degree of freedom is $\chi^2_{0.05\ (2)} = 5.99$.

Decision

$H = 1.54 < \chi^2_{0.05\ (2)} = 5.99$, accept H_0 at 0.05 level of significance.

Conclusion

The samples come from identical population.

Test Statistics

F = 0.9

Critical value

The critical value is $p = P(F_r > 0.9) = 0.9$ for $n = 4$ and $k = 4$

Decision

$P = 0.9 > \alpha = 0.05$, accept H_0 at 5% level of significance.

Conclusion: Birth rate is constant over all four seasons.