

NON-PARAMETRIC STATISTICAL TESTS

Non-parametric tests are normally based on **ranks** of the data samples, and test hypotheses relating to **quantiles** of the probability distribution representing the population from which the data are drawn. Specifically, tests concern the **population median**, η , where

$$\Pr[\text{Observation} \leq \eta] = \frac{1}{2}$$

The **sample median**, x_{MED} , is the mid-point of the sorted sample; if the data $x_1, \dots, x_n$ are sorted into **ascending** order, then

$$x_{\text{MED}} = \begin{cases} x_m & n \text{ odd}, n = 2m + 1 \\ \frac{x_m + x_{m+1}}{2} & n \text{ even}, n = 2m \end{cases}$$

INDEPENDENT SAMPLES

ONE SAMPLE TEST FOR MEDIAN: THE SIGN TEST

For a single sample of size n , to test the hypothesis $\eta = \eta_0$ for some specified value η_0 we use the **Sign Test**. The test statistic S depends on the alternative hypothesis, H_a .

(a) For **one-sided** tests, to test

$$\begin{aligned} H_0 &: \eta = \eta_0 \\ H_a &: \eta > \eta_0 \end{aligned}$$

we define test statistic S by

$$S = \text{Number of observations } \mathbf{greater\ than} \eta_0$$

whereas to test

$$\begin{aligned} H_0 &: \eta = \eta_0 \\ H_a &: \eta < \eta_0 \end{aligned}$$

we define S by

$$S = \text{Number of observations } \mathbf{less\ than} \eta_0$$

If H_0 is **true**, it follows that

$$S \sim \text{Binomial}\left(n, \frac{1}{2}\right)$$

The p-value is defined by

$$p = \Pr[X \geq S]$$

where $X \sim \text{Binomial}(n, 1/2)$. The rejection region for significance level α is defined implicitly by the rule

Reject H_0 if $\alpha \geq p$.

The Binomial distribution is tabulated in McClave and Sincich.

(b) For a **two-sided** test,

$$H_0 : \eta = \eta_0$$

$$H_a : \eta \neq \eta_0$$

we define the test statistic by

$$S = \max\{S_1, S_2\}$$

where S_1 and S_2 are the counts of the number of observations less than, and greater than, η_0 respectively. The p-value is defined by

$$p = 2 \Pr[X \geq S]$$

where $X \sim \text{Binomial}(n, 1/2)$.

Notes :

- 1 The only assumption behind the test is that the data are drawn independently from a continuous distribution.
- 2 If any data are equal to η_0 , we **discard** them before carrying out the test.
- 3 **Large sample approximation.** If n is large (say $n \geq 30$), and $X \sim \text{Binomial}(n, 1/2)$, then it can be shown that

$$X \approx \text{Normal}(np, np(1-p))$$

Thus for the sign test, where $p = 1/2$, we can use the test statistic

$$Z = \frac{S - \frac{n}{2}}{\sqrt{n \times \frac{1}{2} \times \frac{1}{2}}} = \frac{S - \frac{n}{2}}{\sqrt{n} \times \frac{1}{2}}$$

and note that if H_0 is true,

$$Z \approx \text{Normal}(0, 1).$$

so that the test at $\alpha = 0.05$ uses the following critical values

$$H_a : \eta > \eta_0 \quad \text{then} \quad C_R = 1.645$$

$$H_a : \eta < \eta_0 \quad \text{then} \quad C_R = -1.645$$

$$H_a : \eta \neq \eta_0 \quad \text{then} \quad C_R = \pm 1.960$$

- 4 For the large sample approximation, it is common to make a **continuity correction**, where we replace S by $S - 1/2$ in the definition of Z

$$Z = \frac{\left(S - \frac{1}{2}\right) - \frac{n}{2}}{\sqrt{n} \times \frac{1}{2}}$$

Tables of the standard Normal distribution are given in McClave and Sincich.

TWO SAMPLE TEST: THE MANN-WHITNEY-WILCOXON TEST

For a two **independent** samples of size n_1 and n_2 , to test the hypothesis of **equal population medians**

$$H_0 : \eta_1 = \eta_2$$

we use the **Wilcoxon Rank Sum Test**, or an equivalent test, the **Mann-Whitney U Test**; we refer to this as the

Mann-Whitney-Wilcoxon (MWW) Test

By convention it is usual to formulate the test statistic in terms of the **smaller** sample size. Without loss of generality, we label the samples such that $n_1 > n_2$. The test is based on the **sum of the ranks** for the data from sample 2.

EXAMPLE $n_1 = 4, n_2 = 3$ yields the following ranked data

SAMPLE 1	0.31	0.48	1.02	3.11			
SAMPLE 2	0.16	0.20	1.97				
SAMPLE	2	2	1	1	1	2	1
	0.16	0.20	0.31	0.48	1.02	1.97	3.11
RANK	1	2	3	4	5	6	7

Thus the rank sum for sample 1 is

$$R_1 = 3 + 4 + 5 + 7 = 19$$

and the rank sum for sample 2 is

$$R_2 = 1 + 2 + 6 = 9.$$

Two related test statistics can be used

- **Wilcoxon Rank Sum Statistic**

$$W = R_2$$

- **Mann-Whitney U Statistic**

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

We again consider three alternative hypotheses:

$$H_a : \eta_1 < \eta_2$$

$$H_a : \eta_1 > \eta_2$$

$$H_a : \eta_1 = \eta_2$$

and define the rejection region separately in each case.

- **Large Sample Test** If $n_2 \geq 10$, a large sample test based on the Z statistic

$$Z = \frac{U - \frac{n_1 n_2}{2}}{\sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}}$$

can be used. Under the hypothesis $H_0 : \eta_1 = \eta_2$,

$$Z \approx \text{Normal}(0, 1)$$

so that the test at $\alpha = 0.05$ uses the following critical values

$$H_a : \eta_1 > \eta_2 \quad \text{then} \quad C_R = -1.645$$

$$H_a : \eta_1 < \eta_2 \quad \text{then} \quad C_R = 1.645$$

$$H_a : \eta_1 \neq \eta_2 \quad \text{then} \quad C_R = \pm 1.960$$

- **Small Sample Test** If $n_1 < 10$, an **exact** but more complicated test can be used. The test statistic is R_2 (the sum of the ranks for sample 2). The null distribution under the hypothesis $H_0 : \eta_1 = \eta_2$ can be computed, but it is complicated.

The table in McClave and Sincich gives the critical values (T_L and T_U) that determine the rejection region for different n_1 and n_2 values up to 10.

- **One-sided tests:** These are tests at the $\alpha = 0.025$ significance level.

$$H_a : \eta_1 > \eta_2 \quad \text{Rejection Region is} \quad R_2 \leq T_L$$

$$H_a : \eta_1 < \eta_2 \quad \text{Rejection Region is} \quad R_2 \geq T_U$$

- **Two-sided tests:** This is a test at the $\alpha = 0.05$ significance level.

$$H_a : \eta_1 \neq \eta_2 \quad \text{Rejection Region is} \quad R_2 \leq T_L \text{ or } R_2 \geq T_U$$

Notes:

- The only assumption is are needed for the test to be valid is that the samples are independently drawn from two continuous distributions.
- The sum of the ranks across **both** samples is

$$R_1 + R_2 = \frac{(n_1 + n_2)(n_1 + n_2 + 1)}{2}$$

- If there are **ties** (equal values) in the data, then the rank values are replaced by **average** rank values.

DATA VALUE	0.16	0.20	0.31	0.31	0.48	1.97	3.11
ACTUAL RANK	1	2	3	3	5	6	7
AVERAGE RANK	1	2	3.5	3.5	5	6	7

EXAMPLE Sign Test: Water Content Example

The following data are measurements of percentage water content of soil samples collected by two experimenters. We wish to test the hypothesis

$$H_0 : \eta = 9.0$$

for each experiment.

Experimenter 1:	$n = 10$	5.5	6.0	6.5	7.6	7.6	7.7	8.0	8.2	9.1	15.1
Experimenter 2:	$n = 20$	5.6	6.1	6.3	6.3	6.5	6.6	7.0	7.5	7.9	8.0
		8.0	8.1	8.1	8.2	8.4	8.5	8.7	9.4	14.3	26.0

To perform the test, we need tables of the Binomial distribution with $p = 1/2$. The individual probabilities are given by the formula

$$\Pr[X = x] = \binom{n}{x} p^x (1-p)^{n-x} = \binom{n}{x} \frac{1}{2^n} = \frac{n!}{x!(n-x)!} \frac{1}{2^n} \quad x = 0, 1, \dots, n$$

We test at the $\alpha = 0.05$ level. For the first experiment, with $n = 10$:

- If

$$H_a : \eta > 9.0$$

the test statistic is

$$S = \text{Number of observations **greater than** 9} \quad \therefore \quad S = 2$$

and the p-value is

$$p = \Pr[X \geq 2] = 1 - \Pr[X < 2] = 1 - \Pr[X = 0] - \Pr[X = 1] = 0.9893$$

so we **do not** reject H_0 in favour of this H_a .

- If

$$H_a : \eta < 9.0$$

the test statistic is

$$S = \text{Number of observations **less than** 9} \quad \therefore \quad S = 8$$

and the p-value is

$$p = \Pr[X \geq 8] = \Pr[X = 8] + \Pr[X = 9] + \Pr[X = 10] = 0.0547$$

so we **do not** reject H_0 in favour of this H_a .

- If

$$H_a : \eta \neq 9.0$$

the test statistic is

$$S = \max\{S_1, S_2\} = \max\{2, 8\} = 8$$

and the p-value is

$$p = 2\Pr[X \geq 8] = 2(\Pr[X = 8] + \Pr[X = 9] + \Pr[X = 10]) = 0.1094$$

so we **do not** reject H_0 in favour of this H_a .

For the second experiment, with $n = 20$:

- If $H_a : \eta > 9.0$, the test statistic is $S = 3$. The p-value is

$$p = \Pr[X \geq 3] = 1 - \Pr[X < 3] = 1 - \Pr[X = 0] - \Pr[X = 1] - \Pr[X = 2] = 0.9998.$$

so we **do not** reject H_0 in favour of this H_a .

- If $H_a : \eta < 9.0$, the test statistic $S = 17$. The p-value is

$$p = \Pr[X \geq 17] = \Pr[X = 17] + \Pr[X = 18] + \Pr[X = 19] + \Pr[X = 20] = 0.0013.$$

so we **do** reject H_0 in favour of this H_a .

- If $H_a : \eta \neq 9.0$, the test statistic is $S = \max\{S_1, S_2\} = \max\{3, 17\} = 17$. The p-value is

$$p = 2\Pr[X \geq 17] = 2(\Pr[X = 17] + \Pr[X = 18] + \Pr[X = 19] + \Pr[X = 20]) = 0.0026.$$

so we **do** reject H_0 in favour of this H_a .

In R, the sign test can be carried out using basic functions:

```
> #Sign test: Test statistic
> S<-sum(Sign1$h2o>9)
> #First p-value
> 1-pbinom(S-1,nrow(Sign1),0.5)
+ [1] 0.989258

> #Second p-value
> pbinom(S,nrow(Sign1),0.5)
+ [1] 0.0546875

> #Third p-value
> 2*(1-pbinom(max(S,nrow(Sign1)-S)-1,nrow(Sign1),0.5))
+ [1] 0.109375
```

```
> eta<-9
> binom.test(sum(Sign1$h2o>eta),nrow(Sign1),p=0.5,
+           alternative="greater")
+
+ Exact binomial test
+
+ data: sum(Sign1$h2o > eta) and nrow(Sign1)
+ number of successes = 2, number of trials = 10, p-value = 0.989
+ alternative hypothesis: true probability of success is greater than 0.5
+ 95 percent confidence interval:
+ 0.0367714 1.0000000
+ sample estimates:
+ probability of success
+ 0.2
```

```

> binom.test(sum(Sign1$h2o>eta),nrow(Sign1),p=0.5,
+           alternative="less")
+
+ Exact binomial test
+
+ data:  sum(Sign1$h2o > eta) and nrow(Sign1)
+ number of successes = 2, number of trials = 10, p-value = 0.0547
+ alternative hypothesis: true probability of success is less than 0.5
+ 95 percent confidence interval:
+  0.000000 0.506901
+ sample estimates:
+ probability of success
+                0.2

```

```

> binom.test(sum(Sign1$h2o>eta),nrow(Sign1),p=0.5,
+           alternative="two.sided")
+
+ Exact binomial test
+
+ data:  sum(Sign1$h2o > eta) and nrow(Sign1)
+ number of successes = 2, number of trials = 10, p-value = 0.109
+ alternative hypothesis: true probability of success is not equal to 0.5
+ 95 percent confidence interval:
+  0.0252107 0.5560955
+ sample estimates:
+ probability of success
+                0.2

```

Using the Mann-Whitney-Wilcoxon test:

```

> #Mann-Whitney-Wilcoxon test
> Sample1<-c(0.31,0.48,1.02,3.11)
> Sample2<-c(0.16,0.20,1.97)
> Y<-c(Sample1,Sample2)
> Samp<-rep(c(1,2),c(4,3))
> rY<-rank(Y)
> (R1<-sum(rY[Samp==1]))
+ [1] 19
> (R2<-sum(rY[Samp==2]))
+ [1] 9

```

In R, the function `wilcox.test` carries out the test.

```

> wilcox.test(Y~Samp)
+
+ Wilcoxon rank sum exact test
+
+ data:  Y by Samp

```

```

+ W = 9, p-value = 0.4
+ alternative hypothesis: true location shift is not equal to 0
> wilcox.test(Y~Samp, alternative="less")
+
+ Wilcoxon rank sum exact test
+
+ data: Y by Samp
+ W = 9, p-value = 0.886
+ alternative hypothesis: true location shift is less than 0

```

Using the coin library function wilcox_test

```

> coin::wilcox_test(Y~as.factor(Samp)) #Asymptotic approximation
+
+ Asymptotic Wilcoxon-Mann-Whitney Test
+
+ data: Y by as.factor(Samp) (1, 2)
+ Z = 1.061, p-value = 0.289
+ alternative hypothesis: true mu is not equal to 0
> coin::wilcox_test(Y~as.factor(Samp), distribution="exact") #Exact null
+
+ Exact Wilcoxon-Mann-Whitney Test
+
+ data: Y by as.factor(Samp) (1, 2)
+ Z = 1.061, p-value = 0.4
+ alternative hypothesis: true mu is not equal to 0

```

We can also use a permutation procedure:

```

> #Approximate null using permutation Monte Carlo
> coin::wilcox_test(Y~as.factor(Samp), distribution="approximate")
+
+ Approximative Wilcoxon-Mann-Whitney Test
+
+ data: Y by as.factor(Samp) (1, 2)
+ Z = 1.061, p-value = 0.397
+ alternative hypothesis: true mu is not equal to 0

```

EXAMPLE Mann-Whitney-Wilcoxon Test: Low Birthweight Example

The birthweights (in grammes) of babies born to two groups of mothers A and B are displayed below: Thus $n_1 = 9, n_2 = 8$. From this sample (which has ties, so we need to use average ranks),

Group A:	$n = 9$	2164	2600	2184	2080	1820	2496	2184	2080	2184
Group B:	$n = 8$	2576	3224	2704	2912	2444	3120	2912	3848	

so that the two statistics are $R_1 = 48$, $R_2 = 105$, and

$$\text{Wilcoxon } W = R_2 = 105$$

$$\text{Mann-Whitney } U = R_2 - \frac{n_2(n_2 + 1)}{2} = 105 - 36 = 69$$

- For the **small sample** test, from tables in McClave and Sincich, we find

$$T_L = 51 \quad T_U = 93$$

Thus $W > 93$, so we

Do not reject H_0 against $H_a : \eta_1 > \eta_2$ as $W = R_2 > T_L$

Reject H_0 against $H_a : \eta_1 < \eta_2$ as $W = R_2 > T_U$

Reject H_0 against $H_a : \eta_1 \neq \eta_2$ as $W = R_2 > T_U$

The one-sided tests are carried out at $\alpha = 0.025$, the two-sided test is carried out at $\alpha = 0.05$.

- For the **large sample** test, we find

$$Z = \frac{U - \frac{n_1 n_2}{2}}{\sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}} = 3.175$$

Thus we

Do not reject H_0 against $H_a : \eta_1 > \eta_2$ as $Z > C_R = -1.645$

Reject H_0 against $H_a : \eta_1 < \eta_2$ as $Z > C_R = 1.645$

Reject H_0 against $H_a : \eta_1 \neq \eta_2$ as $Z > C_{R_2} = 1.960$

All tests are carried out at $\alpha = 0.05$.

```
> R1<-sum(rank(Birthweight$BW)[Birthweight$gp==1])
> R2<-sum(rank(Birthweight$BW)[Birthweight$gp==2])
> n1<-sum(Birthweight$gp==1); n2<-sum(Birthweight$gp==2)
> c(R1, R2, n1)

+ [1] 58 95 10

> W<-R2
> U<-W-0.5*n2*(n2+1)
> Z<-(U-0.5*n1*n2)/sqrt(n1*n2*(n1+n2+1)/12)
```

Using the Mann-Whitney-Wilcoxon tests:

```
> wilcox.test(BW~Group, data=Birthweight)

+ Warning in wilcox.test.default(x = DATA[[1L]], y = DATA[[2L]], ...):
cannot compute exact p-value with ties
```

```

+
+ Wilcoxon rank sum test with continuity correction
+
+ data: BW by Group
+ W = 3, p-value = 0.00203
+ alternative hypothesis: true location shift is not equal to 0
> wilcox.test(BW~Group,data=Birthweight,alternative="greater")
+ Warning in wilcox.test.default(x = DATA[[1L]], y = DATA[[2L]], ...):
+ cannot compute exact p-value with ties
+
+ Wilcoxon rank sum test with continuity correction
+
+ data: BW by Group
+ W = 3, p-value = 0.999
+ alternative hypothesis: true location shift is greater than 0
> coin::wilcox_test(BW~Group,data=Birthweight,
+ dist="exact",ties.method="mid-ranks")
+
+ Exact Wilcoxon-Mann-Whitney Test
+
+ data: BW by Group (A, B)
+ Z = -3.134, p-value = 0.000668
+ alternative hypothesis: true mu is not equal to 0
> coin::wilcox_test(BW~Group,data=Birthweight,
+ dist="exact",alternative="greater")
+
+ Exact Wilcoxon-Mann-Whitney Test
+
+ data: BW by Group (A, B)
+ Z = -3.134, p-value = 1
+ alternative hypothesis: true mu is greater than 0
> coin::wilcox_test(BW~Group,data=Birthweight,
+ dist="exact",alternative="less")
+
+ Exact Wilcoxon-Mann-Whitney Test
+
+ data: BW by Group (A, B)
+ Z = -3.134, p-value = 0.000309
+ alternative hypothesis: true mu is less than 0

```

EXAMPLE Mann-Whitney-Wilcoxon Test: Treadmill Test Example

The treadmill stress test times (in seconds) of two groups of patients (disease group and healthy controls) are displayed below:

Disease :	$n = 10$	864	636	638	708	786	600	1320	750	594	750
Healthy :	$n = 8$	1014	684	810	990	840	978	1002	1110		

Thus $n_1 = 10, n_2 = 8$. For this sample (which has ties, so we need to use average ranks)

$$R_1 = 70 \quad R_2 = 101$$

so that the two statistics are

$$\text{Wilcoxon } W = R_2 = 101$$

$$\text{Mann-Whitney } U = R_2 - \frac{n_2(n_2 + 1)}{2} = 101 - 36 = 65$$

- For the **small sample** test, from tables in McClave and Sincich, we find

$$T_L = 54 \quad T_U = 98$$

Thus $W > 98$, so we

Do not reject H_0 against $H_a : \eta_1 > \eta_2$ as $W = R_2 > T_L$

Reject H_0 against $H_a : \eta_1 < \eta_2$ as $W = R_2 > T_U$

Reject H_0 against $H_a : \eta_1 \neq \eta_2$ as $W = R_2 > T_U$

One-sided tests are carried out at $\alpha = 0.025$; two-sided test is carried out at $\alpha = 0.05$.

- For the **large sample** test, we find

$$Z = \frac{U - \frac{n_1 n_2}{2}}{\sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}} = 2.221$$

Thus we

Do not reject H_0 against $H_a : \eta_1 > \eta_2$ as $Z > C_R = -1.645$

Reject H_0 against $H_a : \eta_1 < \eta_2$ as $Z > C_R = 1.645$

Reject H_0 against $H_a : \eta_1 \neq \eta_2$ as $Z > C_{R_2} = 1.960$

All tests are carried out at $\alpha = 0.05$.

```
> wilcox.test(Time~Group, data=Treadmill)
+ Warning in wilcox.test.default(x = DATA[[1L]], y = DATA[[2L]], ...):
+ cannot compute exact p-value with ties
+
+ Wilcoxon rank sum test with continuity correction
+
+ data: Time by Group
+ W = 65, p-value = 0.0294
+ alternative hypothesis: true location shift is not equal to 0
```

```

> wilcox.test(Time~Group,data=Treadmill,alternative="greater")

+ Warning in wilcox.test.default(x = DATA[[1L]], y = DATA[[2L]], ...):
cannot compute exact p-value with ties

+
+ Wilcoxon rank sum test with continuity correction
+
+ data: Time by Group
+ W = 65, p-value = 0.0147
+ alternative hypothesis: true location shift is greater than 0

> wilcox.test(Time~Group,data=Treadmill,alternative="less")

+ Warning in wilcox.test.default(x = DATA[[1L]], y = DATA[[2L]], ...):
cannot compute exact p-value with ties

+
+ Wilcoxon rank sum test with continuity correction
+
+ data: Time by Group
+ W = 65, p-value = 0.988
+ alternative hypothesis: true location shift is less than 0

> coin::wilcox_test(Time~as.factor(Group),data=Treadmill,
+ dist="exact", ties.method="mid-ranks")

+
+ Exact Wilcoxon-Mann-Whitney Test
+
+ data: Time by as.factor(Group) (1, 2)
+ Z = 2.222, p-value = 0.0249
+ alternative hypothesis: true mu is not equal to 0

> coin::wilcox_test(Time~as.factor(Group),data=Treadmill,
+ dist="exact", alternative="greater")

+
+ Exact Wilcoxon-Mann-Whitney Test
+
+ data: Time by as.factor(Group) (1, 2)
+ Z = 2.222, p-value = 0.0125
+ alternative hypothesis: true mu is greater than 0

> coin::wilcox_test(Time~as.factor(Group),data=Treadmill,
+ dist="exact", alternative="less")

+
+ Exact Wilcoxon-Mann-Whitney Test
+
+ data: Time by as.factor(Group) (1, 2)
+ Z = 2.222, p-value = 0.989
+ alternative hypothesis: true mu is less than 0

```

DEPENDENT SAMPLES

TWO DEPENDENT SAMPLES: WILCOXON SIGNED RANK TEST

Data collected from the same experimental units are in general **dependent**. For example, if data are collected on two occasions (time 1 and time 2, or before and after treatment) from the same n individuals, then the resulting data samples $(y_{11}, \dots, y_{n1})$ and $(y_{12}, \dots, y_{n2})$ are dependent. Such data are often referred to as **paired**. We wish to test whether there is a significant change across the two measurements.

For a **parametric** test, we typically assume that the within-individual differences

$$x_i = y_{i1} - y_{i2} \quad i = 1, \dots, n$$

are **Normally** distributed, and test the hypothesis that the mean difference μ is zero

$$H_0 : \mu = 0$$

using a one-sample Z -test (σ known) or T -test (σ unknown), with statistic

$$z = \frac{\bar{x}}{\sigma/\sqrt{n}} \quad \text{or} \quad t = \frac{\bar{x}}{s/\sqrt{n}}$$

distributed as $\text{Normal}(0, 1)$ or t_{n-1} respectively.

For a **non-parametric** test, we can use the **Wilcoxon Signed Rank** test:

1. Compute the within-individual differences

$$x_i = y_{i1} - y_{i2} \quad i = 1, \dots, n$$

If any $x_i = 0$, then that data point is discarded and the sample size adjusted.

2. Sort the **absolute values** $s_1, \dots, s_n$ of $x_1, x_2, \dots, x_n$ into **ascending** order, and assign ranks 1 up to n . If there are ties, assign **average** ranks.
3. Form the two rank sums T_+ and T_- , where

$$T_+ = \text{Sum of ranks for those } x_i > 0 \quad T_- = \text{Sum of ranks for those } x_i < 0$$

The test statistic is a function of these rank sums. Heuristically, if the statistic T_+ is large and T_- is small, this implies that the experimental units where $y_{i1} > y_{i2}$ have a **larger** (in magnitude) difference than those where $y_{i1} < y_{i2}$. This indicates an overall **decrease** between the first and second measurements. Conversely, if the statistic T_- is large and T_+ is small, this implies that the experimental units where $y_{i2} > y_{i1}$ have a **larger** (in magnitude) difference than those where $y_{i2} < y_{i1}$. This indicates an overall **increase** between the first and second measurements.

We test the null hypothesis

$$H_0 : \text{No change between first and second measurements}$$

against the three alternative hypotheses

- (1) H_a : Significant **decrease** between first and second measurements
- (2) H_a : Significant **increase** between first and second measurements
- (3) H_a : Significant **change** between first and second measurements

To test H_0 vs (1), we perform a one-sided test using the statistic T_- ; the critical value in the test is denoted T_0 , and is determined by the table in McClave and Sincich:

If $T_- \leq T_0$, we **reject** H_0 in favour of H_a (1)

To test H_0 vs (2), we perform a one-sided test using the statistic T_+ ; the critical value is T_0 and

If $T_+ \leq T_0$, we **reject** H_0 in favour of H_a (2)

To test H_0 vs (3), we perform a two-sided test using the statistic $T = \min\{T_-, T_+\}$; the critical value is T_0

If $T \leq T_0$, we **reject** H_0 in favour of H_a (3)

Notes :

- 1 The only assumption behind the test is that the difference data x_i are drawn independently from a continuous distribution.
- 2 **Large Sample Test:** For $n \geq 25$, we can use a large sample version of the test based on T_+ , and the Z statistic

$$Z = \frac{T_+ - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}}$$

If H_0 is **true**, then $Z \sim \text{Normal}(0, 1)$, so that the test at $\alpha = 0.05$ uses the following critical values

For H_a (1) use $C_R = 1.645$

For H_a (2) use $C_R = -1.645$

For H_a (3) use $C_R = \pm 1.960$

EXAMPLE Haemodialysis Data

The following data are measurements of the heparin cofactor II (HCII) to plasma protein ratios in a group of patients at baseline and five months after haemodialysis.

Patient	Before	After				
	y_{i1}	y_{i2}	x_i	s_i	Rank	Ave. Rank
1	2.11	2.15	-0.04	0.04	3	3.5
2	1.85	2.11	-0.26	0.26	10	10.0
3	1.82	1.93	-0.11	0.11	8	8.0
4	1.75	1.83	-0.08	0.08	6	6.0
5	1.54	1.90	-0.36	0.36	11	11.0
6	1.52	1.56	-0.04	0.04	3	3.5
7	1.49	1.44	0.05	0.05	5	5.0
8	1.44	1.43	0.01	0.01	1	1.5
9	1.38	1.28	0.10	0.10	7	7.0
10	1.30	1.30	0.00	0.00	-	-
11	1.20	1.21	-0.01	0.01	1	1.5
12	1.19	1.30	-0.11	0.11	9	9.0
						$T_+ = 13.5$
						$T_- = 52.5$

OMIT

From the textbook table, for $n = 12 - 1 = 11$, we find that the $\alpha = 0.025/0.05$ (one/two-sided) significance level critical value is $T_0 = 11$. Thus using T_+ , we **cannot reject** either of the null

hypotheses (2) and (3), as $T_+ > T_0$. Note that $Z = -1.734$, so if the approximation was valid, we would be able to reject (2) at $\alpha = 0.05$.

```
> #Haemodialysis data
> wilcox.test(Haemo$after[-10], Haemo$before[-10],
+   paired=T)

+ Warning in wilcox.test.default(Haemo$after[-10], Haemo$before[-10], paired
= T): cannot compute exact p-value with ties

+
+ Wilcoxon signed rank test with continuity correction
+
+ data: Haemo$after[-10] and Haemo$before[-10]
+ V = 52.5, p-value = 0.0908
+ alternative hypothesis: true location shift is not equal to 0

> wilcox.test(Haemo$after[-10], Haemo$before[-10],
+   paired=T, alternative="greater")

+ Warning in wilcox.test.default(Haemo$after[-10], Haemo$before[-10], paired
= T, : cannot compute exact p-value with ties

+
+ Wilcoxon signed rank test with continuity correction
+
+ data: Haemo$after[-10] and Haemo$before[-10]
+ V = 52.5, p-value = 0.0454
+ alternative hypothesis: true location shift is greater than 0

> #Direct calculation
> xi<-Haemo$after-Haemo$before
> xi<-xi[xi!=0]
> si<-abs(xi)
> n<-length(si)
>
> Tplus<-sum(rank(si)[xi > 0])
> Tminus<-sum(rank(si)[xi < 0])
>
> (Z<-(Tplus-0.25*n*(n+1))/sqrt(n*(n + 1)*(2*n + 1)/24))

+ [1] 1.73376

> pnorm(Z)

+ [1] 0.95852
```

EXAMPLE Asthma data:

```
> PEFRAsthma #PEFRAsthma data
+ Patient before after
+ 1      1      312    300
+ 2      2      342    201
+ 3      3      340    232
+ 4      4      388    312
+ 5      5      296    220
+ 6      6      254    256
+ 7      7      391    328
+ 8      8      402    330
+ 9      9      290    231

> PEFRAsthma$before-PEFRAsthma$after #Differences
+ [1] 12 141 108 76 76 -2 63 72 59

> wilcox.test(PEFRAsthma$after,PEFRAsthma$before,paired=T)
+ Warning in wilcox.test.default(PEFRAsthma$after, PEFRAsthma$before, paired
= T): cannot compute exact p-value with ties
+
+ Wilcoxon signed rank test with continuity correction
+
+ data: PEFRAsthma$after and PEFRAsthma$before
+ V = 1, p-value = 0.0128
+ alternative hypothesis: true location shift is not equal to 0
> wilcox.test(PEFRAsthma$after,PEFRAsthma$before,paired=T,
+ alternative="less")
+ Warning in wilcox.test.default(PEFRAsthma$after, PEFRAsthma$before, paired
= T, : cannot compute exact p-value with ties
+
+ Wilcoxon signed rank test with continuity correction
+
+ data: PEFRAsthma$after and PEFRAsthma$before
+ V = 1, p-value = 0.00639
+ alternative hypothesis: true location shift is less than 0
> #Direct calculation
> xi<-PEFRAsthma$after-PEFRAsthma$before
> xi<-xi[xi!=0]
> si<-abs(xi); n<-length(si)
>
> Tplus<-sum(rank(si)[xi > 0])
> Tminus<-sum(rank(si)[xi < 0])
>
> (Z<-(Tplus-0.25*n*(n+1))/sqrt(n*(n + 1)*(2*n + 1)/24))
+ [1] -2.5471
> pnorm(Z)
+ [1] 0.00543111
```

THREE OR MORE INDEPENDENT SAMPLES

We now seek non-parametric tests that can be used for multiple independent samples, such as those found in the Completely Randomized Design (CRD) and Randomized Block Design (RBD) described in the ANOVA section. The non-parametric equivalents of the F tests for these two designs are

- The **Kruskal-Wallis H test** for a Completely Randomized Design
- **Friedman's test** for a Randomized Block Design

KRUSKAL-WALLIS TEST

In a CRD, we have k independent groups, corresponding to k different treatments, with sample sizes $n_1, \dots, n_k$. Let $n = n_1 + \dots + n_k$. To compute the test statistic, H , we

1. Pool the data, sort them into ascending order, and assign ranks. If there are ties in the data, then average ranks are used.
2. For $j = 1, \dots, k$, compute the rank sum R_j

$$R_j = \text{Sum of ranks for data from sample } j.$$

To test the hypothesis

$$\begin{aligned} H_0 &: \text{No difference between the population distributions of the } k \text{ groups} \\ H_a &: \text{At least two population distributions different} \end{aligned}$$

the test statistic is

$$H = \frac{12}{n(n+1)} \sum_{j=1}^k \frac{R_j^2}{n_j} - 3(n+1)$$

If H_0 is **true**, then for large n ,

$$H \approx \chi_{k-1}^2.$$

Notes :

- 1 The test assumes that the k samples are independently drawn from continuous populations.
- 2 For the approximation to be valid, there should be at least **five** observations in each sample, and the number of ties should be small.

EXAMPLE Mucociliary efficiency data

The data are measures of mucociliary efficiency from the rate of removal of dust in normal subjects (Group 1), subjects with obstructive airway disease (Group 2), and subjects with asbestosis (Group 3).

Reference: Myles Hollander, M and Douglas A. Wolfe (1973), *Nonparametric statistical inference*, New York: John Wiley & Sons. pp115-120.

Group	1	1	1	1	1	2	2	2	2	3	3	3	3	3
y	2.9	3.0	2.5	2.6	3.2	3.8	2.7	4.0	2.4	2.8	3.4	3.7	2.2	2.0
Rank	8	9	4	5	10	13	6	14	3	7	11	12	2	1

Hence $R_1 = 36$, $R_2 = 36$ and $R_3 = 33$, and the test statistic $H = 0.7714$. To complete the test, we compare with the $\alpha = 0.05$ quantile of the $\chi^2_{k-1} = \chi^2_2$ distribution. We have

$$\chi^2_{0.05,2} = 5.99 > H \quad \therefore \quad \text{No evidence to reject } H_0$$

and a p-value of $p = 0.680$.

FRIEDMAN TEST

In a RBD, we have k treatment groups, and a blocking factor. For example, we might have k repeated measurements on the same b experimental units, and $n = bk$ observations in total. To compute the test statistic, F_r , we proceed as follows.

1. **Within each block separately**, sort the k data values into ascending order, and assign ranks. If there are ties in the data, then average ranks are used.
2. For $j = 1, \dots, k$, compute the rank sum R_j

$$R_j = \text{Sum of ranks for data from treatment } j.$$

To test the hypothesis

H_0 : No difference between the population distributions of the k treatment groups

H_a : At least two population distributions different

the test statistic is

$$F_r = \frac{12}{bk(k+1)} \sum_{j=1}^k R_j^2 - 3b(k+1)$$

If H_0 is **true**, then for large n ,

$$F_r \approx \chi^2_{k-1}$$

Notes :

- 1 The test assumes that the data are drawn independently from continuous populations, with random assignment of treatments within blocks.
- 2 For the approximation to be valid, it is recommended that b or k is at least five, and the number of ties should be small.

EXAMPLE Skin potential under hypnosis

A study was conducted to investigate whether hypnosis has the same effect on skin potential for four different emotions. Eight subjects were asked to display fear, joy, sadness and calmness under hypnosis, and the resulting skin potential (measured in millivolts) was recorded for each emotion. Thus in this experiment, $b = 8$ and $k = 4$.

Subject	Fear		Joy		Sadness		Calmness	
	y	Rank	y	Rank	y	Rank	y	Rank
1	23.1	4	22.7	3	22.5	1	22.6	2
2	57.6	4	53.2	2	53.7	3	53.1	1
3	10.5	3	9.7	2	10.8	4	8.3	1
4	23.6	4	19.6	3	21.1	2	21.6	1
5	11.9	1	13.8	4	13.7	3	13.3	2
6	54.6	4	47.1	3	39.2	2	37.0	1
7	21.0	4	13.6	1	13.7	2	14.8	3
8	20.3	3	23.6	4	16.3	2	14.8	1
Rank Sum	27		20		19		14	

Thus the within-treatment rank sums are $R_1 = 27, R_2 = 20, R_3 = 19$ and $R_4 = 14$ and thus $F_r = 6.45$. To complete the test, we compare with the $\alpha = 0.05$ quantile of the

$$\chi_{k-1}^2 = \chi_3^2$$

distribution. We have

$$\chi_{0.05,3}^2 = 7.81 > F_r \quad \therefore \quad \text{No evidence to reject } H_0$$

and a p-value of $p = 0.092$.

```
> #Mucociliary data
> #Mean ranks
> xtabs(rank(y) ~ group, data=Muco) / xtabs(~group, data=Muco)

+ group
+           Healthy Obstructive airway disease
+           7.2           9.0
+           Asbestosis
+           6.6

> kruskal.test(y ~ group, data=Muco)

+
+  Kruskal-Wallis rank sum test
+
+ data:  y by group
+ Kruskal-Wallis chi-squared = 0.7714, df = 2, p-value = 0.68

> #MemoryTask
> #Mean ranks
> xtabs(rank(Memory) ~ Task, data=Memory) / xtabs(~Task, data=Memory)
```

```

+ Task
+   Counting      Rhyming      Adjective      Imagery Intentional
+      12.95       13.10       31.50       36.60       33.35

> kruskal.test(Memory~Task, data=Memory)

+
+   Kruskal-Wallis rank sum test
+
+ data:  Memory by Task
+ Kruskal-Wallis chi-squared = 25.38, df = 4, p-value = 4.23e-05

```

```

> #Skin potential data (Hypnosis)
>
> friedman.test(SkinResponse~EmoCode | Subject, data=Hypnosis)

+
+   Friedman rank sum test
+
+ data:  SkinResponse and EmoCode and Subject
+ Friedman chi-squared = 6.45, df = 3, p-value = 0.0917

> #Soil
> friedman.test(sulphur~solvent | soil, data=Soil)

+
+   Friedman rank sum test
+
+ data:  sulphur and solvent and soil
+ Friedman chi-squared = 1.08, df = 3, p-value = 0.782

```