Note: For the MWW test

- **Textbook convention :** Label the samples so that $n_1 > n_2$ (i.e. sample 1 is the one with the larger sample size)
- **SPSS convention :** Label the samples such that

$$x_{\text{MED}_1} < x_{\text{MED}_2}$$

(i.e. sample 1 is the one with the smaller median) and only test

$$H_0 : \eta_1 = \eta_2$$

Other two sample tests are available:

- Kolmogorov-Smirnov Test
- Moses Extreme Reactions Test
- Wald-Wolfowitz Runs Test

None make distributional assumptions, all perform best when the sample size is large.

Non-parametric Statistics
Comparing Two Populations
Comparing Two Dependent Samples
Comparing Three or More Populations

3.4 Comparing Two Dependent Samples

Suppose we have repeat measurements on the same experimental units.

In this case, the **within-subject** data are **dependent**; we have pairing of observations.

We can use the

Wilcoxon Signed Rank Test

SEE HANDOUT

Non-parametric Statistics
Comparing Two Populations
Comparing Two Dependent Samples
Comparing Three or More Populations

3.4 Comparing Three or More Populations

We now seek non-parametric equivalents to ANOVA useful for different designs. We study tests for

(a) the **Completely Randomized Design** (CRD)

(b) the **Randomized Block Design** (RBD)

For (a) we use the

Kruskal-Wallis Test

and for (b) we use the

Friedman Test.

SEE HANDOUT

Non-parametric Statistics

Two Dependent Samples and Multiple Independent Samples

1. 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 $(y_{11}, \ldots, y_{n1})$ and $(y_{12}, \ldots, y_{n2})$ 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

$$x_i = y_{i1} - y_{i2} \qquad i = 1, \ldots, n$$

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

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

distributed as Normal$(0, 1)$ or Student$(n - 1)$ respectively.

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

1. Compute the within-individual differences

$$x_i = y_{i1} - y_{i2} \qquad i = 1, \ldots, n$$

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

2. Sort the **absolute values** $s_1, \ldots, s_n$ of $x_1, x_2, \ldots, x_n$ 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

$$\begin{aligned} T_+ &= \text{Sum of ranks for those } x_i > 0 \\ T_- &= \text{Sum of ranks for those } x_i < 0 \end{aligned}$$

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

$$(1)\ H_a \quad : \quad \text{Significant } \textbf{decrease} \text{ between first and second measurements}$$

$$(2)\ H_a \quad : \quad \text{Significant } \textbf{increase} \text{ between first and second measurements}$$

$$(3)\ H_a \quad : \quad \text{Significant } \textbf{change} \text{ between first and second measurements}$$

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 on p. 839 of McClave and Sincich:

$$\text{If } T_- \leq T_0, \text{ we } \textbf{reject } H_0 \text{ in favour of } H_a \text{ (1)}$$

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

$$\text{If } T_+ \leq T_0, \text{ we } \textbf{reject } H_0 \text{ in favour of } H_a \text{ (2)}$$

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 and

$$\text{If } T \leq T_0, \text{ we } \textbf{reject } H_0 \text{ in favour of } H_a \text{ (3)}$$

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 $Z \dot{\sim} \text{Normal}(0, 1)$, 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 1: 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.

Reference: Toulon, P *et al*. (1987) Antithrombin III and heparin cofactor II in patients with chronic renal failure undergoing regular hemodialysis, *Thrombosis and Haemostasis*, **3**;57(3): pp263-8.

Patient	Before y_{i1}	After 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	**OMIT**	**OMIT**
11	1.20	1.21	-0.01	0.01	1	1.5
12	1.19	1.30	-0.11	0.11	9	9.0
						$T_+ = 13.5$
						$T_- = 52.5$

From the table on p 839, 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 $\alpha = 0.05$.

2. Three or more independent samples: The Kruskal-Wallis and Friedman Tests

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 Fisher-F tests ANOVA for these two designs are

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

2.1 Kruskal-Wallis Test

In a CRD, we have k independent groups, corresponding to k different treatments, with sample sizes $n_1, \ldots, n_k$. Let $n = n_1 + \cdots + n_k$. 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 $j = 1, \ldots, k$, compute the rank sum R_j

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

To test the hypothesis

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

H_a : At least two population distributions different

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,

$$H \dot{\sim} \text{Chisquared}(k-1).$$

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 2: Mucociliary efficiency data

The following 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

Chisquared$(k-1)$ = Chisquared(2) distribution. We have

$$\text{Chisq}_{0.05}(2) = 5.99 > H \qquad \therefore \qquad \textbf{No evidence to reject } H_0$$

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

2.2 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 $j = 1, \ldots, k$, compute the rank sum R_j

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

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,

$$F_r \dot{\sim} \text{Chisq}(k-1)$$

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 3: 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 $k = 4$.

	Fear		Joy		Sadness		Calmness	
Subject	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

$$R_1 = 27 \qquad R_2 = 20 \qquad R_3 = 19 \qquad R_4 = 14$$

and thus

$$F_r = 6.45$$

To complete the test, we compare with the $\alpha = 0.05$ quantile of the Chisquared$(k-1) =$ Chisquared(3) distribution. We have

$$\text{Chisq}_{0.05}(3) = 7.81 > F_r \qquad \therefore \qquad \textbf{No evidence to reject } H_0$$

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