Data Science 1: Non-Parametric Tests

Ann Maharaj

Non−Parametric Tests

1. Test for the Median of One Population
   - Wilcoxon Signed-Ranks Test for the Median
2. Tests for Equality of Two Population Medians
   - Mann-Whitney U Test for Two Independent Populations
   - Wilcoxon Paired Signed Ranks Test for Related Populations
3. Tests for Equality of More Than Two Population Medians
   - Kruskal-Wallis Test: One-Way Analysis of Variance by Ranks
   - Friedman Test: Two-Way Analysis of Variance by Ranks

Wilcoxon Signed-Ranks Test for the Median

- The *t*-test for the mean of a population is based on the assumption that the variable of interest is normally distributed.
  - Hence the variable is quantitative and continuous.
- If the sample is of a reasonable size, the presence of extreme values could contribute to the departure from normality.
  - While removing of extreme values could reduce the departure from normality, it is not always a good idea to do so since in many cases extreme values are part of the data set and should be included in further analysis.

Wilcoxon Signed-Ranks Test for the Median

- Hence, if the assumption of normality is not valid, a non-parametric alternative should be considered.
- Another reason to use a non-parametric test is when sample size is too small to check for normality of the underlying distribution.
- The Wilcoxon Signed-Ranks Test for the Median is one such non-parametric alternative to the t-test.

Wilcoxon Signed-Ranks Test for the Median

- **Assumption**
  The sample available for analysis is a random sample of independent measurements from a symmetric population with unknown median, M.

- **Data**
  The variable of interest is quantitative but could also be qualitative on the ordinal scale.

- **Hypotheses**
  $H_0 : M = m_0 \quad H_1 : M \neq m_0$
  $H_0 : M = m_0 \quad H_1 : M > m_0$
  $H_0 : M = m_0 \quad H_1 : M < m_0$

Wilcoxon Signed-Ranks Test for the Median

Given a variable of interest X and a sample of size n, the test is implemented as follows:

- Determine $d_i = x_i - m_0$, for $i = 1, 2, \ldots, n$ where m_0 is the hypothesised median.
- Rank the differences d_i from smallest to largest without regard to their signs.
- If a difference is negative, the rank is given a negative sign, if it is positive, it is given a positive sign.
- The sum of the positive ranks W_+, and the sum of the negative ranks W_- are obtained. It can be shown that
  $W_+ = [n(n+1)/2] - W_-$.
- If any observation x_i is equal m_0, it is removed and the sample size is reduced accordingly.
- When ties occur between ranks, the ranks are averaged over the values.

Wilcoxon Signed-Ranks Test for the Median

- If H_0 is true and if the assumptions are met:
  - The probability of observing a positive difference $d_i = x_i - m_0$ of a given magnitude is equal to observing a negative difference of the same magnitude.
  - In repeated sampling, $E(W_+) = E(W_-)$.
- For a given sample, we do not expect $W_+ = W_-$, but if H_0 is true , we do not expect a big difference in their values.
- If H_0 is true we would expect the distribution of the differences to be approximately symmetric around zero and the distribution of positives and negatives to be distributed at random among the ranks. Under this assumption, it is possible to work out the exact probability of every possible outcome for W.
- Test Statistic: Wilcoxon Rank Sum Statistic, $W = min(W_+, W_-)$;
- P-value: Knowing the distribution of W under H_0, the exact p-value can be determined, viz.,
  p-value = P(W $<$ numerical value of the test statistic) for a one-tailed alternative.
  p-value = 2P(W $<$ numerical value of the test statistic) for the two-tailed alternative.

Wilcoxon Signed Ranks Test for the Median

- For large sample sizes, it can be shown that W is approximately normally distributed with mean and standard deviation:

$$\mu_W = \frac{n(1+n)}{4}$$

$$\sigma_W = \sqrt{\frac{n(1+n)(2n+1)}{24}}.$$

- Hence the standardised W follows a standard normal distribution:

$$Z = \frac{W - \mu_W}{\sigma_W} \sim N(0,1).$$

- In this case, the normal probability distribution is used to obtain the p-value of the test.

Example

- Suppose that the median house price in an upmarket suburb of a capital city in Australia is 2 million AUD . Over the last two weeks a real estate agent sold 5 houses at the following prices in millions of AUD in this suburb:

  2.15, 1.13, 2.25, 2.50, 2.30

- If this is considered to be random sample of houses sold in this suburb over the last two weeks, test if the median house price in this suburb has been exceeded.
- Why is the use of t-test for the population mean not appropriate in this case?

Example: Wilcoxon Signed Ranks Test for the Median

M: Population median

1. Hypotheses:
   $H_0 : M = 2$
   $H_1 : M > 2$
2. Level of significance: $\alpha = 0.05$
3. Test statistic: $V = 10$
4. P-Value: p-value $= 0.313$
5. Decision Rule: Reject H_0 if p-value < 0.05
6. Conclusion: Since p-value > 0.05, do not reject H_0 at the 5% level and indeed any reasonable level of significance.

```
> #Wilcoxon signed rank test for median
> wilcox.test(price, alternative="greater", mu=2,conf.int=TRUE)

        Wilcoxon signed rank exact test

data:  price
V = 10, p-value = 0.3125
alternative hypothesis: true location is greater than 2
95 percent confidence interval:
 1.13  Inf
sample estimates:
(pseudo)median
        2.225
```

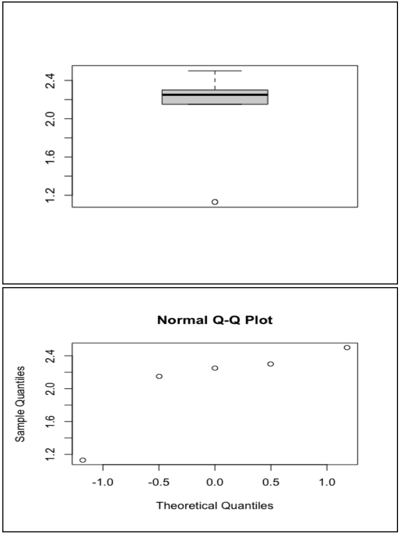
There is not sufficient evidence to conclude that the median house price has exceeded 2 million AUD .

Example: Wilcoxon Signed Ranks Test for the Median

```
        Shapiro-Wilk normality test

data:  price
W = 0.77264, p-value = 0.04759
```

The t-test for the population mean is not appropriate because:

- As can be observed from the outcome of the Shapiro-Wilk test and the Normal Q-Q plot, the assumption of normality is not valid.
- Furthermore, the data set is too small for a reasonably assessment of the normality assumption and if we remove the out outlier observed in the boxplot, it will be reduced even further.



Mann-Whitney U Test for Independent Populations

- The Mann-Whitney U test is the non-parametric counterpart of the t-test used to compare the means of two independent populations.
- **Assumptions**

  1) Random samples are selected from two independent populations of interest.

  2) The distribution functions of the two populations differ only with respect to location.
- **Data**

  The variables of interest, X and Y are quantitative but could also be qualitative on the ordinal scale.
- **Hypotheses**

  $H_0 : M_X = M_Y \quad H_1 : M_X \neq M_Y$
  Rejecting H_0 implies a significant shift in central location to the left or to the right.

  $H_0 : M_X = M_Y \quad H_1 : M_X > M_Y$
  Rejecting H_0 implies a significant shift in central location to the right.

  $H_0 : M_X = M_Y \quad H_1 : M_X < M_Y$
  Rejecting H_0 implies a significant shift in central location to the left.

Mann-Whitney U Test for Independent Populations

- Computation of the test begins by arbitrarily designating two samples as Group 1 and Group 2.
- The data from the two groups are combined into one group with each data value retaining a group identifier of its original group.
- The pooled values are ranked from 1 to n, with the smallest value being assigned a rank of 1.
- When ties occur between ranks, the ranks are averaged over the values.
- The sum of ranks of values from Group 1 is computed and designated as W_1, and sum of ranks of values from Group 2 is computed and designated as W_2 .

Mann-Whitney U Test for Independent Populations

- The next step is to calculate a test statistic for each of W_1 and W_2 as follow:

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - W_1$$

$$U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - W_2$$

- **Test Statistic**

$$U = min(U_1, U_2)$$

- The probability associated with distribution of U is determined exactly so that the p-value of the test can be determined.
  p-value $= \mathrm{P}(U <$ numerical value of the test statistic$)$ for a one-tailed alternative.
  p-value $= 2\mathrm{P}(U <$ numerical value of the test statistic$)$ for the two-tailed alternative.

Mann-Whitney U Test for two Independent Populations

- For large sample sizes, it can be shown that U is approximately normally distributed with mean and standard deviation

$$\mu_U = \frac{n_1 n_2}{2}$$

$$\sigma_U = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}$$

- Hence the standardised U follows a standard normal distribution.

$$Z = \frac{U - \mu_U}{\sigma_U} \sim N(0, 1)$$

- In this case, the normal probability distribution is used to obtain the p-value of the test.

Example

- Suppose that a random sample of 7 health-sector support workers along with a random sample of 8 education-sector support workers are taken in a particular city and information is obtained about their hourly wages. The data is shown below and in the file hwage.csv.

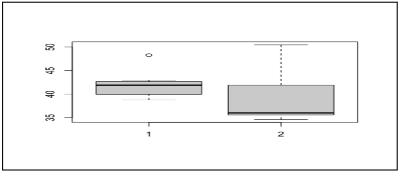
Health Sector	Education Sector
40.10	46.19
39.80	36.20
42.36	35.60
38.75	37.60
41.90	35.52
42.96	35.80
48.25	34.60
	50.45

- Using the Mann-Whitney U test, can it be concluded that there is a difference between the hourly wages of health-sector and education-sector support workers. Test at the at the 5% level of significance .
- Explain why this test is preferable to the t-test for two independent populations.

Example: Mann-Whitney U Test

M_X: Median hourly wages in Health Sector, M_Y: Median hourly wages in Education Sector

1. Hypotheses:
   $H_0 : M_X = M_Y$
   $H_1 : M_X \neq M_Y$
2. Level of significance: $\alpha = 0.05$
3. Test statistic: W $= 43$
4. P-Value: p-value $= 0.094$
5. Decision Rule: Reject H_0 if p-value < 0.05
6. Conclusion: Since p-value > 0.05, do not reject H_0 at the 5% level of significance.

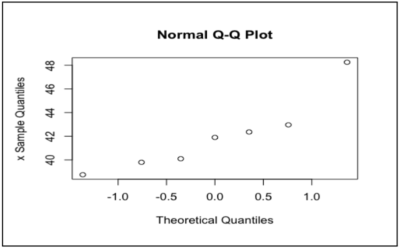


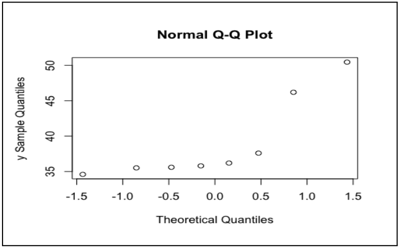

```
> #Mann Whitney U test
> #(Wilcoxon test) for medians of 2 independent populations
> wilcox.test(x,y, paired=FALSE, conf.int=TRUE)

        Wilcoxon rank sum exact test

data:  x and y
W = 43, p-value = 0.09386
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval:
 -3.83  7.30
sample estimates:
difference in location
                   4.3
```

There is not sufficient evidence to conclude that there is a difference in the hourly wages of health-sector and education-sector support workers.

However, if we test at the 10% level of significance, there is sufficient evidence to conclude that there is a difference in the hourly wages of health-sector and education-sector support workers.

Example: Mann-Whitney *U* Test






```
> shapiro.test(x)

        Shapiro-Wilk normality test

data:  x
W = 0.87568, p-value = 0.2079

> shapiro.test(y)

        Shapiro-Wilk normality test

data:  y
W = 0.72356, p-value = 0.004149
```

The Mann U test for equality of two population medians is preferable to the t-test for equality of two population means because it can be observed from the outcome of the Shapiro-Wilk test and the Normal Q-Q plot for one of the variables, the assumption of normality which is a requirement of the t-test, is not valid.

Wilcoxon Paired Signed Ranks Test for Related Populations

- The Wilcoxon Paired Signed Ranks Test is the non-parametric alternative to the t-test for two related populations.
- An example of a related samples scenario is: before and after studies in which measures are taken on the same person or item under two different conditions.
- **Assumptions**

  1) A random sample of items or individuals is selected but because they are subjected to two different conditions, they can be regarded as being samples from two related populations.

  2) The underlying distributions are symmetrical.
- **Data**

  The variables of interest are quantitative but could also be qualitative on the ordinal scale.

Wilcoxon Paired Signed Ranks Test for Related Populations

- The Wilcoxon test utilises the differences of the scores of the two matched groups in a manner similar to that of the t-test for two related samples of size n.
- After the difference scores have been computed, the differences are ranked regardless of whether a difference is positive or negative.
- The values are ranked from smallest to largest, with a rank of 1 assigned to the smallest difference.
- If a difference is negative, the rank is given a negative sign. If a difference is positive, the rank is given a positive sign.

Wilcoxon Paired Signed Ranks Test for Related Populations

- The sum of the positive ranks and the sum of the negative ranks are obtained.
- Zero differences representing ties between the scores from the two groups are ignored and the value of the sample size n is reduced accordingly.
- When ties occur between ranks, the ranks are averaged over the values.
- The sum of the positive ranks W_+, and the sum of the negative ranks W_- are obtained. it can be shown that $W_+ = [n(n+1)/2] - W_-$.

Wilcoxon Paired Signed Ranks Test for Related Populations

- **Hypotheses**

  $H_0 : M_d = 0 \quad H_1 : M_d \neq 0$
  $H_0 : M_d = 0 \quad H_1 : M_d > 0$
  $H_0 : M_d = 0 \quad H_1 : M_d < 0$

  M_d is the median of the difference.

- Test Statistic: Wilcoxon Paired Rank Sum Statistic,
  $W = min(W_+, W_-)$

- P-value: Knowing the distribution of W under H_0, the exact p-value can be determined, viz.,
  p-value = P(W < numerical value of the test statistic) for a one-tailed alternative.
  p-value = 2P(W < numerical value of the test statistic) for the two-tailed alternative.

Wilcoxon Paired Signed Ranks Test for Related Populations

- For large sample sizes, it can be shown that W is approximately normally distributed with mean and standard deviation:

$$\mu_W = \frac{n(1+n)}{4}$$

$$\sigma_W = \sqrt{\frac{n(1+n)(2n+1)}{24}}.$$

- Hence the standardised W follows a standard normal distribution:

$$Z = \frac{W - \mu_W}{\sigma_W} \sim N(0,1).$$

- In this case, the normal probability distribution is used to obtain the p-value of the test.

Example

- The table below shows the hours of relief provided by two analgesic drugs taken by a sample of 12 patients suffering from arthritis, during two different time periods.
- Using the Wilcoxon paired signed rank procedure test, is there any evidence that Drug B provides longer relief than Drug A? Test at the 5% level of significance.
- Explain why this test is preferable to the t-test for paired samples.

Data file: drugs.csv

Case	Drug A	Drug B
1	2.0	3.5
2	3.6	5.7
3	2.6	2.9
4	2.6	2.4
5	7.3	9.9
6	3.4	3.3
7	14.9	16.7
8	6.6	6.0
9	2.3	3.8
10	2.0	4.0
11	6.8	9.1
12	8.5	20.9

Example: Wilcoxon Paired Signed Ranks Test

d: Difference between hours of relief for Drugs B and A.
M_d: Population median of the differences

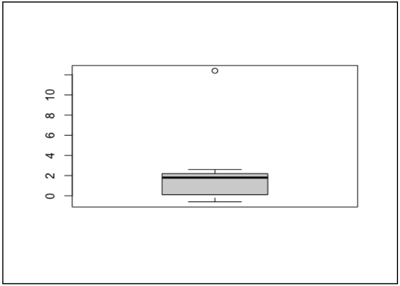
1. Hypotheses:
   $H_0 : M_d = 0$
   $H_1 : M_d > 0$
2. Level of significance: $\alpha = 0.05$
3. Test statistic: $V = 59$
4. P-Value: p-value $= 0.009$
5. Decision Rule: Reject H_0 if p-value < 0.05
6. Conclusion: Since p-value < 0.05, reject H_0 at the 5% level and indeed any reasonable level of significance.

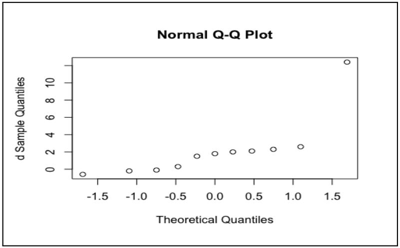
```
> wilcox.test(x1,y1, paired=TRUE, alternative = "greater", conf.int=TRUE)

        Wilcoxon signed rank exact test

data:  x1 and y1
V = 59, p-value = 0.009277
alternative hypothesis: true location shift is greater than 0
95 percent confidence interval:
 0.7 Inf
sample estimates:
(pseudo)median
         1.375
```

There is sufficient evidence to conclude that Drug B provides longer relief than Drug A .

Example: Wilcoxon Paired Signed Ranks Test

```
        Shapiro-Wilk normality test

data:  d1
W = 0.64194, p-value = 8.524e-05

> #Wilcocox test for for medians of 2 related populations
> wilcox.test(x1,y1, paired=TRUE, conf.int=TRUE)
```





The Wilcoxon Paired Signed Ranks Test is preferable to the Paired t-test because it can be observed from the outcome of the Shapiro-Wilk test and the Normal Q-Q plot for differences, the assumption of normality which is a requirement for the paired t-test, is not valid.

Kruskal-Wallis Test: One-Way Analysis of Variance by Ranks

- The non-parametric alternative to the **Completely Randomised Design**, i.e., the one-way analysis of variance, is the **Kruskal-Wallis** procedure.
- Like the one-way ANOVA, this test is used to determine if two or more samples ($c \geq 2$) come from the same or different populations.
- The Kruskal-Wallis Test is used when the underlying assumptions of the one-way ANOVA cannot be met.
- It can also be applied to ordinal data.

Kruskal-Wallis Test: One-Way Analysis of Variance by Ranks

- The objective is to determine if we may conclude from sample evidence that there is a difference among treatment effects, i.e., effects under different conditions.
- If the treatments do not differ in their effects, the median response of a population of observations being subjected to a given treatment will be the same as the median of the observations subjected to any one of the other treatments under study,
- When there are just two samples, the Kruskal-Wallis test is equivalent to the Mann-Whitney U test.

Kruskal-Wallis Test: One-Way Analysis of Variance by Ranks

- **Assumption**
  Random samples are selected from c independent populations of interest.

- **Data**
  The variables of interest are quantitative but could also be qualitative on the ordinal scale.

- **Hypotheses**
  $H_0 : M_1 = M_2 = ... = M_c$
  H_1: Not all Mj are equal, $j = 1, 2, ..., c$

Kruskal-Wallis Test: One-Way Analysis of Variance by Ranks

- The process of computing the Kruskall-Wallis test statistic begins with ranking the data in all the groups together as though were from one group.
- The smallest value is given the rank 1.
- For ties, each value is given the average ranks for the tied values.
- Unlike the one-way ANOVA, in which the raw data are analysed, the Kruskall-Wallis test analyses the ranks of the data.

Kruskal-Wallis Test: One-Way Analysis of Variance by Ranks

- **Test Statistic**

$$K = \frac{12}{n(n-1)}\left(\sum_{j=1}^{c}\frac{T_j^2}{n_j}\right) - 3(n+1)$$

c = number of groups
n = total sample size
T_j = sum of the ranks in a group j
n_j= sample size for group j.

- The test statistic K has an approximate chi-square distribution with c-1 degrees of freedom.

- When the total sample size n is very small, exact probabilities associated with K are determined.

Example

- Four tour guides employed at a state-based tourist attraction of historical interest conduct tours of exactly 20 people each during peak seasons.
- A researcher conducted a study to compare the tour guides with respect to efficiency. For the tour guides labelled A, B, C and D, the researcher selected a random sample of 6 tours during the busiest summer month, and the time in minutes for each was recorded. The results are shown below and in the file tour.csv.

A	B	C	D
40	36	38	35
41	33	38	44
33	38	39	41
30	36	36	30
39	37	38	45
31	39	37	44

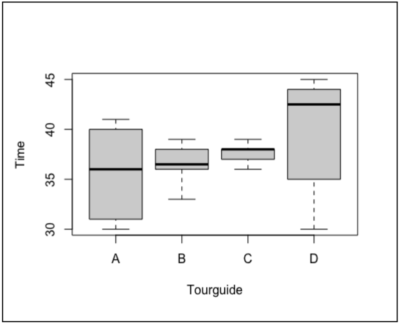
- Using the Kruskal-Wallis test, can the researcher conclude on the basis of these data that the four tour guides differ with respect to the median time it takes them to conduct a tour. Use the 5% level of significance.
- Explain why this test is preferable to the completely randomised design test procedure.

Example: Kruskal-Wallis Test

M_1, M_2, M_3, M_4 are the median tour times of tour guides A, B, C, D, respectively.

1. Hypotheses:
   $H_0 : M_1 = M_2 = M_3 = M_4$
   H_1: Not all M_j are equal, $j = 1, 2, 3, 4$
2. Level of significance: $\alpha = 0.05$
3. Test statistic: Kruskal-Wallis chi-squared $= 2.5505$
4. P-Value: p-value $= 0.466$
5. Decision Rule: Reject H_0 if p-value < 0.05
6. Conclusion: Since p-value > 0.05, do not reject H_0 at the 5% level and indeed any reasonable level of significance.

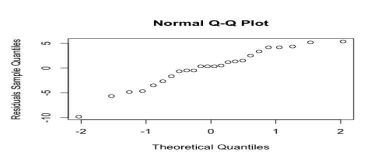
There is not sufficient evidence to conclude that four tour guides differ with respect to the median time it takes them to conduct their tours.



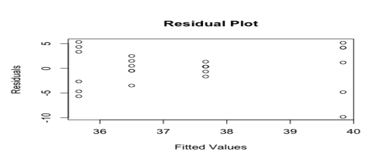

```
> kruskal.test(Time~Tourguide)

        Kruskal-Wallis rank sum test

data:  Time by Tourguide
Kruskal-Wallis chi-squared = 2.5505, df = 3, p-value = 0.4662
```

Example: Kruskal-Wallis Test

```
        Shapiro-Wilk normality test

data:  res
W = 0.94908, p-value = 0.2588
```




```
Levene's Test for Homogeneity of Variance (center = median)
      Df F value  Pr(>F)
group  3  3.7409 0.02776 *
      20
```



- The requirement for one-way ANOVA test for the completely randomised design is normality and constant variances.
- While the assumption of normality is valid as observed from the outcome of the Shapiro-Wilk test and the Normal Q-Q plot, the residual plot and the outcome of Levene's test for constant variances reveals that the assumption of constant variances is not valid.
- Hence the Kruskal-Wallis Test is preferable to the Completely Randomised Design Procedure .

Friedman Test: Two-Way Analysis of Variance by Ranks

- The non-parametric alternative to the **Randomised Block design,** i.e., two-way analysis of variance where we have a blocking variable, is the **Friedman procedure**.
- The Friedman test is used when the underlying assumptions of the two-way ANOVA cannot be met.
- It can also be applied to ordinal data.

Friedman Test: Two-Way Analysis of Variance by Ranks

- The objective is to determine if we may conclude from sample evidence that there is a difference among treatment effects, i.e., effects under different conditions.
- If the treatments do not differ in their effects, the median response of a population of observations being subjected to a given treatment will be the same as the median of the observations subjected to any one of the other treatments under study, after the effect of the blocking variable has been removed.

Friedman Test: Two-Way Analysis of Variance by Ranks

- **Assumptions**
  1) The data for analysis consists of responses of r blocks to c independent conditions or treatments.
  2) A random sample of items or individuals (blocks) is selected but because they are subjected to different conditions, they can be regarded as being samples from c related populations.
  3) There is no interaction between the blocks and treatments.

- **Data**
  The variable of interest is quantitative but could also be qualitative on the ordinal scale.

Friedman Test: Two-Way Analysis of Variance by Ranks

- **Hypotheses**
  $H_0 : M_1 = M_2 = ... = M_c$
  H_1: Not all Mj are equal, $j = 1, 2, ..., c$

- The process of computing the Friedman test statistic begins with ranking the observations in each block from smallest to largest so that each block contains a set of c ranks.

- If the null hypothesis is true, and all treatments have identical effects, the rank that appears in a particular column is merely a matter of chance, i.e., the ranks should be randomly distributed over the columns (treatments) in each row (block).

- If the null hypothesis is false, we expect a lack of randomness in this distribution.

Friedman Test: Two-Way Analysis of Variance by Ranks

- The next step in calculating the test statistic is to obtain the sums of the ranks in each column.
- If the null hypothesis is true, we expect the sum of the ranks to be fairly close in size, that we can attribute differences to chance.
- If the null hypothesis is false, we expect to see at least one difference in size from at least one other sum that we do not attribute the difference to chance alone.

Friedman Test: Two-Way Analysis of Variance by Ranks

- **Test Statistic**

$$Fr = \frac{12}{rc(c+1)} \sum_{j=1}^{c} \left[R_j - \frac{r(c+1)}{2} \right]^2$$

c = number of treatments
r = number of blocks
R_j = sum of the ranks for treatment j

- The test statistic Fr has an approximate chi-square distribution with c-1 degrees of freedom.

- When the sample size is very small, exact probabilities associated with Fr are determined.

Example

- Suppose that a fast-food chain wants to evaluate the service at three of its restaurants.
- The customer service for the chain hires six investigators, with varied experience in food service evaluation, as raters.
- Raters randomly select the order of three restaurants to provide their ratings on a scale of 1 to 10 which are shown below and in the file ratings.cvs

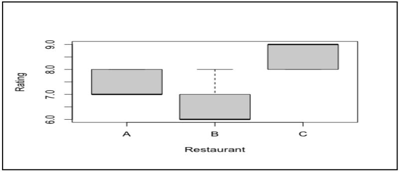
	Restaurant		
Rater	A	B	C
1	7	6	8
2	7	7	8
3	7	9	6
4	8	6	9
5	8	6	9
6	7	8	9

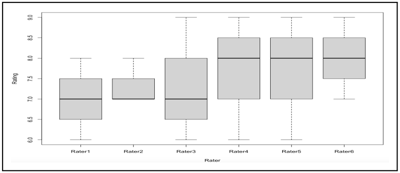
- Using the Friedman test, can it be concluded at the 5% level of significance that there is a difference between the median ratings of the three restaurants.
- Explain why the randomised block design test procedure will not be appropriate in this case.

Example: Friedman Test

M_1, M_2, M_3 are the median ratings of Restaurants A, B, C, respectively.

1. Hypotheses:
   $H_0 : M_1 = M_2 = M_3$
   H_1: Not all M_j are equal, $j = 1, 2, 3$
2. Level of significance: $\alpha = 0.05$
3. Test statistic: Friedman chi-squared = 10.174
4. P-Value: p-value = 0.006
5. Decision Rule: Reject H_0 if p-value < 0.05
6. Conclusion: Since p-value < 0.05 reject H_0 at the 5% level and indeed any reasonable level of significance. There is sufficient evidence to conclude that the ratings of the 3 restaurants differ.






```
> #Friedman Test: Two-Way Analysis of Variance by Ranks
> friedman.test(Rating, Restaurant, Rater)

        Friedman rank sum test

data:  Rating, Restaurant and Rater
Friedman chi-squared = 10.174, df = 2, p-value = 0.006177
```

Example: Friedman Test

- To perform the two-way ANOVA test for the randomised block design procedure, the date must be quantitative and the assumptions of normality and constant variances should be valid.
- In this case the data is qualitative on the ordinal scale, i.e., the ratings are ordered labels.
- Hence, the Friedman Test which is also applicable to ordinal-scaled data, is appropriate in this case, whereas the Randomised Block Design Procedure is not.