1. Water samples were taken at three different locations in a river to determine whether the quantity of dissolved oxygen, a measure of water pollution, varied from one location to another. Location one was selected above an industrial plant; location two was adjacent to the industrial water discharge for the plant; and location three was slightly downriver in midstream. Five water specimens were randomly selected at each location. The data are shown in the accompanying table (the greater the pollution, the lower will be the dissolved oxygen readings).

<table>
<thead>
<tr>
<th>Location 1</th>
<th>Location 2</th>
<th>Location 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.9</td>
<td>4.8</td>
<td>6.0</td>
</tr>
<tr>
<td>6.1</td>
<td>5.0</td>
<td>6.1</td>
</tr>
<tr>
<td>6.3</td>
<td>4.3</td>
<td>5.8</td>
</tr>
<tr>
<td>6.1</td>
<td>4.7</td>
<td>5.6</td>
</tr>
<tr>
<td>6.0</td>
<td>5.1</td>
<td>5.7</td>
</tr>
</tbody>
</table>

\[ \sum \sum x_{ij} = 83.5, \quad \sum \sum x_{ij}^2 = 470.25. \]

(a) Test whether the data provide sufficient evidence to indicate a difference in mean dissolved content for the three locations by using the Kruskal-Wallis test (Hint: \[ \sum \sum r_{ij}^2 = 1237.5. \])

(b) Test whether there is a difference in mean dissolved content between Location 1 and Location 3 by using individual comparison.

Solution:

(a) Let \( \mu_1, \mu_2 \) and \( \mu_3 \) denote the mean quantity of dissolved oxygen at the three locations respectively. The hypothesis to be tested is

\[ H : \mu_1 = \mu_2 = \mu_3 \quad vs \quad H_A : \text{Not all the } \mu_j \text{'s are equal}. \]

The rank matrix of the data is given as follows (rank all the data from all groups together):

<table>
<thead>
<tr>
<th>Location 1</th>
<th>Location 2</th>
<th>Location 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>9.0</td>
<td>3</td>
<td>10.5</td>
</tr>
<tr>
<td>13.0</td>
<td>4</td>
<td>13.0</td>
</tr>
<tr>
<td>15.0</td>
<td>1</td>
<td>8.0</td>
</tr>
<tr>
<td>13.0</td>
<td>2</td>
<td>6.0</td>
</tr>
<tr>
<td>10.5</td>
<td>5</td>
<td>7.0</td>
</tr>
</tbody>
</table>
From the rank matrix, we obtain that

\[ n_i = 5 \quad 5 \quad 5 \quad N = 15 \quad g = 3 \]
\[ \bar{r}_i = \begin{pmatrix} 12.1 \\ 3.0 \\ 8.9 \end{pmatrix} \quad \bar{r} = 8 \]
\[ \sum n_i (\bar{r}_i - \bar{r})^2 = 5 \sum (\bar{r}_i)^2 - 15(\bar{r})^2 = 213.1 \]
\[ \sum \sum (r_{ij} - \bar{r})^2 = \sum \sum r_{ij}^2 - 15(\bar{r})^2 = 277.5 \]
\[ K = \frac{(15 - 1) \sum n_i (\bar{r}_i - \bar{r})^2}{\sum \sum (r_{ij} - \bar{r})^2} = 10.75099. \]

The corresponding \( p \)-value is

\[ p = P \left( \chi^2_2 \geq K \right) = P \left( \chi^2_2 \geq 10.75099 \right) < 0.01. \]

**Conclusion:** The data is against the null hypothesis.

(b) We have to test

\[ H : \mu_1 = \mu_3 \quad \text{vs} \quad H_A : \mu_1 \neq \mu_3, \]

by using \( t \)-test related to individual comparison.

By the information contained in Question 1, calculations related are given as follows:

\[ n_i = 5 \quad 5 \quad 5 \quad N = 15 \quad g = 3 \]
\[ \bar{x}_i = \begin{pmatrix} 6.08 \\ 4.78 \\ 5.84 \end{pmatrix} \quad \bar{x} = 5.57 \quad CM = 464.8167 \]

\[ \text{Group SS} = \sum_{i=1}^{3} n_i (\bar{x}_i)^2 - CM = 4.785333 \]
\[ \text{Total SS} = \sum \sum x_{ij}^2 - CM = 5.433333 \]
\[ \text{Residual SS} = \text{Total SS} - \text{Group SS} = 0.648 \]
\[ s^2 = \text{Residual SS}/(15 - 3) = 0.054 \]

A observed value of the \( t \)-test is

\[ t_{1,3} = \frac{\bar{x}_1 - \bar{x}_3}{s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = 1.633 \]

The corresponding \( p \)-value is

\[ p = 2 \cdot P(t_{13} \geq 1.633) > 0.01. \]

**Conclusion:** The data is consistent with the null hypothesis.

2. In the case of no ties, show that the Kruskal-Wallis test statistic

\[ K = \frac{12}{N(N+1)} \sum_{j=1}^{g} n_j (\bar{r}_j)^2 - 3(N + 1). \]
Hint: Recall the remarks in Lecture 18 and also note that if there are no ties on the data, then
\[ \sum_{j=1}^{g} \sum_{i=1}^{n_j} r_{ij}^2 = \sum_{i=1}^{N} i^2 = \frac{1}{6} N(N+1)(2N+1); \]

**Solution.** In the case of no ties,
\[ \sum_{j=1}^{g} \sum_{i=1}^{n_j} r_{ij} = \sum_{i=1}^{N} i = N(N+1)/2, \quad \bar{r} = (N+1)/2, \]
\[ \sum_{j=1}^{g} \sum_{i=1}^{n_j} r_{ij}^2 = \sum_{i=1}^{N} i^2 = \frac{1}{6} N(N+1)(2N+1). \]

This, together with a simple calculation (see remark in Lecture 18), implies that
\[ \sum_{j=1}^{g} n_j (\bar{r}_j - \bar{r})^2 = \sum_{j=1}^{g} n_j (\bar{r}_j)^2 - N(\bar{r})^2 \]
\[ = \sum_{j=1}^{g} n_j (\bar{r}_j)^2 - \frac{1}{4} N(N+1)^2, \]
\[ \sum_{j=1}^{g} \sum_{i=1}^{n_j} (r_{ij} - \bar{r})^2 = \sum_{j=1}^{g} \sum_{i=1}^{n_j} r_{ij}^2 - N(\bar{r})^2 \]
\[ = \frac{1}{6} N(N+1)(2N+1) - \frac{1}{4} N(N+1)^2 \]
\[ = \frac{1}{12} N(N+1) [2(2N+1) - 3(N+1)] \]
\[ = \frac{1}{12} N(N+1)(N-1). \]

Now it follows easily that
\[ K = (N - 1) \frac{\sum_{j=1}^{g} n_j (\bar{r}_j - \bar{r})^2}{\sum_{j=1}^{g} \sum_{i=1}^{n_j} (r_{ij} - \bar{r})^2} \]
\[ = (N - 1) \frac{\sum_{j=1}^{g} n_j (\bar{r}_j)^2 - \frac{1}{4} N(N+1)^2}{\frac{1}{12} N(N+1)(N-1)} \]
\[ = \frac{12}{N(N+1)} \sum_{j=1}^{g} n_j (\bar{r}_j)^2 - 3(N+1). \]

3. Assuming \( g = 2 \), show that the Kruskal-Wallis test statistic \( K = \tilde{W}^2 \), where the \( \tilde{W} \) is the standardised Wilcoxon test statistic defined by
\[ \tilde{W} = \frac{n_1 \bar{r}_1 - n_1(N+1)}{\sqrt{n_1 n_2 (N+1)}}. \]

Hint: \( n_1 + n_2 = N \), \( \bar{r} = (n_1 \bar{r}_1 + n_2 \bar{r}_2)/N = (N+1)/2. \)
Solution. When \( g = 2 \), the Wilcoxon test statistic \( W = n_1 \tilde{r}_1 \).

\[
E(W_i) = \frac{N(N+1)}{2} \cdot \frac{1}{N} = \frac{N+1}{2}
\]

\[
\Rightarrow E(W) = \frac{n_1(N+1)}{2}
\]

\[
Var(W_i) = E(W_i^2) - [E(W_i)]^2 = \frac{N(N+1)(2N+1)}{6} \cdot \frac{1}{N} - \left( \frac{N+1}{2} \right)^2
\]

\[
= \frac{(N+1)(2N^2 + 3N - 3)}{12} = \frac{(N+1)(N-1)}{12} = \frac{N^2 - 1}{12}
\]

\[
\Rightarrow Var(W) = Var\left( \sum_{i=1}^{n_1} W_i \right) = \sum_{i=1}^{n_1} Var(W_i) + 2 \sum_{i<j} Cov(W_i, W_j)
\]

\[
= n_1 Var(W_i) + n_1(n_1 - 1) Cov(W_i, W_j)
\]

The equation holds for all \( n_1 \leq N \). In particular, if \( n_1 = N \), we have

\[
W = \sum_{i=1}^{N} W_i = \frac{1}{2} N(N+1) \quad \text{a constant}
\]

\[
\Rightarrow Var(W) = NVar(W_i) + N(N - 1) Cov(W_i, W_j) = 0
\]

\[
\Rightarrow Cov(W_i, W_j) = -\frac{Var(W_i)}{N - 1}
\]

Hence we have

\[
Var(W) = n_1 Var(W_i) - \frac{n_1(n_1 - 1)}{N - 1} Cov(W_i, W_j)
\]

\[
= \frac{n_1(N-1) - n_1(n_1-1)}{N - 1} Var(W_i)
\]

\[
= \frac{n_1(N-n_1)}{N - 1} Var(W_i) = \frac{n_1(N-n_1)}{N - 1} \cdot \frac{N^2 - 1}{12}
\]

\[
= \frac{n_1(N-n_1)}{N - 1} \times \frac{(N+1)(N-1)}{12} = n_1n_2(N+1)
\]

Hence the standardised Wilcoxon test statistic is

\[
\tilde{W} = \frac{W - E(W)}{\sqrt{Var(W)}} = \frac{n_1 \tilde{r}_1 - \frac{n_1(N+1)}{2}}{\sqrt{\frac{n_1n_2(N+1)}{12}}}.
\]

Then with no ties, we have

\[
K = (N - 1) \frac{n_1(\tilde{r}_{1} - \bar{r})^2 + n_2(\tilde{r}_{2} - \bar{r})^2}{\sum_{i=1}^{n_1} (r_{1i} - \bar{r})^2 + \sum_{i=1}^{n_2} (r_{2i} - \bar{r})^2}
\]

\[
= (N - 1) \frac{n_1\tilde{r}_{1}^2 + n_2\tilde{r}_{2}^2 - N\bar{r}^2}{\sum_{i=1}^{n_1} r_{1i}^2 + \sum_{i=1}^{n_2} r_{2i}^2 - N\bar{r}^2}
\]
\[
\begin{align*}
&= (N - 1) \frac{n_1 \bar{r}_1^2 + n_2 \bar{r}_2^2 - (n_1 \bar{r}_1 + n_2 \bar{r}_2)^2}{N(N+1)(2N+1)} - \frac{N}{4}\\
&= (N - 1) \frac{n_1 \bar{r}_1^2 + n_2 \bar{r}_2^2 - n_1^2 \bar{r}_1^2 - n_2^2 \bar{r}_2^2 - 2n_1 n_2 \bar{r}_1 \bar{r}_2}{N(N+1)(N-1)}\\
&= \frac{n_1 n_2 \bar{r}_2^2}{N^2(N+1)} + \frac{n_1 n_2 \bar{r}_2^2}{N^2(N+1)} - 2n_1 n_2 \bar{r}_1 \bar{r}_2\\
&= \frac{n_1^2 n_2^2 (\bar{r}_1 - \bar{r}_2)^2}{n_1 n_2(N+1)}\\
&= \frac{n_1^2 (n_2 \bar{r}_1 - n_2 \bar{r}_2)^2}{n_1 n_2(N+1)}\\
&= \frac{n_1^2 [(N - n_1) \bar{r}_1 - n_2 \bar{r}_2]^2}{n_1 n_2(N+1)}\\
&= \frac{n_1^2 [N \bar{r}_1 - n_1 \bar{r}_1 - n_2 \bar{r}_2]^2}{n_1 n_2(N+1)}\\
&= \frac{n_1^2 [N \bar{r}_1 - N \bar{r}]^2}{n_1 n_2(N+1)}\\
&= \frac{[n_1 \bar{r}_1 - n_1 \frac{N+1}{2}]^2}{n_1 n_2(N+1)}\\
&= \left[ W - E(W) \right]^2 = \bar{W}^2
\end{align*}
\]
1. Let us consider an experimental study of drugs to relieve itching. Five drugs (N3-N7) were compared to a placebo (N2) and no drugs (N1) with 10 volunteer male subjects aged 20-30.

<table>
<thead>
<tr>
<th></th>
<th>N1</th>
<th>N2</th>
<th>N3</th>
<th>N4</th>
<th>N5</th>
<th>N6</th>
<th>N7</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>174</td>
<td>263</td>
<td>105</td>
<td>199</td>
<td>141</td>
<td>108</td>
<td>141</td>
</tr>
<tr>
<td>2</td>
<td>224</td>
<td>213</td>
<td>103</td>
<td>143</td>
<td>168</td>
<td>341</td>
<td>184</td>
</tr>
<tr>
<td>3</td>
<td>260</td>
<td>231</td>
<td>145</td>
<td>113</td>
<td>78</td>
<td>159</td>
<td>125</td>
</tr>
<tr>
<td>4</td>
<td>255</td>
<td>291</td>
<td>103</td>
<td>225</td>
<td>164</td>
<td>135</td>
<td>227</td>
</tr>
<tr>
<td>5</td>
<td>165</td>
<td>168</td>
<td>144</td>
<td>176</td>
<td>127</td>
<td>239</td>
<td>194</td>
</tr>
<tr>
<td>6</td>
<td>237</td>
<td>121</td>
<td>94</td>
<td>144</td>
<td>114</td>
<td>136</td>
<td>155</td>
</tr>
<tr>
<td>7</td>
<td>191</td>
<td>137</td>
<td>35</td>
<td>87</td>
<td>96</td>
<td>140</td>
<td>121</td>
</tr>
<tr>
<td>8</td>
<td>100</td>
<td>102</td>
<td>133</td>
<td>120</td>
<td>222</td>
<td>134</td>
<td>129</td>
</tr>
<tr>
<td>9</td>
<td>115</td>
<td>89</td>
<td>83</td>
<td>100</td>
<td>165</td>
<td>185</td>
<td>79</td>
</tr>
<tr>
<td>10</td>
<td>189</td>
<td>433</td>
<td>237</td>
<td>173</td>
<td>168</td>
<td>188</td>
<td>317</td>
</tr>
</tbody>
</table>

The data is available in

(a) Open the web page, copy down the data and save it as "itch" in your NEDIT.
(b) Read the data "itch" into Splus for analysis by using

    inch <- read.table("itch", header=F, sep=" ")

(c) Set a multiple graph window with 1 row and 2 columns (Using par or Graph)
(d) Obtain side by side boxplot of the data.
(e) Obtain a normal quantile plot of the combined residuals from the groups N1-N7 by using the following steps:
    i. Create a group mean vector, say g.mean, by using apply.
    ii. Create a matrix, say g.mat, by using matrix() in which the matrix matches to the data matrix "itch" and has the same group mean at each column.
    iii. Find the residual matrix of the data and then perform a normal qq-plot of the combined residuals.
(f) Comment on whether or not the data appears to satisfy the assumptions for an analysis of variance.
(g) Perform a One-Way ANOVA on the data, and comment on the results.
(h) Perform a multiple comparison test on the data and find groups for which the means are significantly different at level $\alpha = 0.05$ and $\alpha = 0.10$.

Solution:
(a) Open the web page, copy down the data and save it as "inch" in your NEDIT.

(b) Read the data "inch" into Splus for analysis by using

```
inch <- read.table("inch", header=F, sep=" ")
```

# (c)
par(mfrow = c(1, 2))  # equivalent to Graph(1,2)

# (d) boxplot(inch)

# (e)
> g.mean <- apply(inch, 2, mean)
> g.mat <- matrix(rep(g.mean, 10), nr = 10, byrow = T)
> residual <- inch - g.mat
> residual

```
 V1    V2    V3    V4    V5    V6    V7
 1   -17  58.2  -13.2  51  -3.3  -68.5  -26.2
 2    33  8.2   -15.2   -5  23.7  164.5  16.8
 3    69  26.2   26.8  -35  -66.3  -17.5  -42.2
 4    64  86.2  -15.2  77  19.7  -41.5  59.8
 5  -26  -36.8  25.8  28  -17.3  62.5  26.8
 6    46  -83.8  -24.2   -4  -30.3  -40.5  -12.2
 7    67  26.2  -15.2  77  19.7  -41.5  59.8
 8   -91  -102.8   14.8  -28  77.7  -42.5  -38.2
 9   -76  -115.8  -35.2  -48  20.7   8.5  -88.2
10    -2  228.2  118.8  25  23.7  11.5  149.8
```

> qqnorm(residual)

# (f)
# The boxplot shows the variance during the groups are nearly the same.
# The normal quantile plot of the residuals is fairly linear except a few outliers.
# A normal model is appropriate and so we can use the F-test in One-Way ANOVA.

# (g)

> inch1<-as.matrix(inch)
> inch1<-as.vector(inch1)
> fac<-factor(rep(letters[1:7], c(10,10,10,10,10,10,10)))
> inch1.df<-data.frame(fac, inch1)
> aov.inch<-aov(inch1~fac, inch1.df)

> summary(aov.inch)

```
 Df  Sum Sq Mean Sq  F value Pr(>F)
 7 1131.76
```

7
# The p-value indicates that borderline evidence against the null hypothesis.

# (h)
> inch.mul<-multicomp(aov.inch, focus="fac", method="bon", alpha=0.05)
> inch.mul
> plot(inch.mul)

> inch.mul1<-multicomp(aov.inch, focus="fac", method="bon", alpha=0.10)
> inch.mul
> plot(inch.mul1)

# There are no groups in which the means are significantly different at level $\alpha=0.05$

# There are marginal evidence that the groups N2 and N3 are significantly different at level $\alpha=0.10$.