THE UNIVERSITY OF SYDNEY

STAT2012 STATISTICAL TESTS

Semester 2	**Solution to Tutorial Week 7**	2014

Tutorial questions

1. Let μ_1, μ_2 and μ_3 denote the mean quantity of dissolved oxygen at the three locations respectively. We have $g = 3$ groups, the total sample size $N = 15$ and the sample size for each group $n_i = 5$. The matrix of data and ranks is given as follows (rank all the data from all groups together):

	Location 1		Location 2		Location 3	
j	y_{1j}	r_{1j}	y_{2j}	r_{2j}	y_{3j}	r_{3j}
1	5.9	9.0	4.8	3	6.0	10.5
2	6.1	13.0	5.0	4	6.1	13.0
3	6.3	15.0	4.3	1	5.8	8.0
4	6.1	13.0	4.7	2	5.6	6.0
5	6.0	10.5	5.1	5	5.7	7.0
Mean $\bar{x}$	6.08	12.1	4.78	3	5.84	8.9
Var. s_i^2	0.022		0.097		0.043	

(a) The KW test for the equality of mean quantity of dissolved oxygen at the three locations is

1. **Hypotheses:** $H_0 : \mu_1 = \mu_2 = \mu_3$ vs H_1 : Not all the μ_j's are equal.
2. **Test statistic:**

$$
\begin{array}{lccc}
n_i & 5 & 5 & 5 \qquad \sum_{i=1}^{g}\sum_{j=1}^{n_i} r_{ij}^2 = 9^2 + \cdots + 7^2 = 1237.5 \\
\bar{r}_{i.} & 12.1 & 3.0 & 8.9 \quad \bar{r} = (N+1)/2 = (15+1)/2 = 8
\end{array}
$$

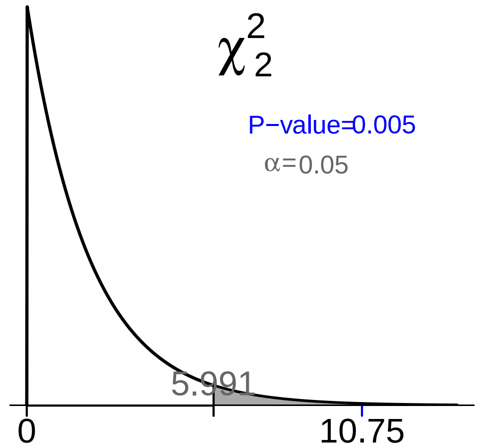
$$SST = \sum_{i=1}^{g} n_i \bar{r}_{i.}^2 - N\bar{r}^2 = 5(12.1^2 + 3^2 + 8.9^2) - 15(8)^2 = 213.1$$

$$MST_o = \frac{1}{N-1}\sum_{i=1}^{g}\sum_{j=1}^{n_i} r_{ij}^2 - N(\bar{r})^2 = \frac{1237.5 - 15(8)^2}{14} = \frac{277.5}{14} = 19.82143$$

$$k_0 = (N-1)\frac{\sum_{i=1}^{g} n_i(\bar{r}_{i.} - \bar{r})^2}{\sum_{i=1}^{g}\sum_{j=1}^{n_i}(r_{ij} - \bar{r})^2} = \frac{SST}{MST_o} = \frac{213.1}{19.82143} = 10.75099.$$

3. **Assumption:** Same distribution of Y_{ij} in each group i. We have $k_0 \sim \chi^2_{g-1}$ under H_0.
4. **P-value:** p-value $= \Pr(\chi^2_2 \geq k_0) = \Pr(\chi^2_2 \geq 10.75099) << 0.01$.

5. **Decision:** Since p-value < 0.05, we reject H_0. There are strong evidence in the data against H_0 that the mean quanitiy of dissolved oxygen at the three locations are equal.



(b) The summary information is given as follow:

n_i	5	5	5	$\sum_{i=1}^{g}\sum_{j=1}^{n_i} y_{ij}^2 = 470.25$
$\bar{y}_{i.}$	6.08	4.78	5.84	$\bar{y} = 5.57$
s_i^2	0.022	0.097	0.043	

The 2 sample t-test for the difference in mean quantity of dissolved oxygen between Location 1 and 3 is

1. **Hypotheses:** $H_0: \ \mu_i = \mu_j \quad \text{vs} \quad H_1: \ \mu_i \neq \mu_j$, etc for each pair (i, j)
2. **Test statistic:** $t_{1,2} = \dfrac{\bar{y}_{1.} - \bar{y}_{2.}}{\sqrt{MSR}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \dfrac{6.08 - 4.78}{0.2324\sqrt{\frac{1}{5} + \frac{1}{5}}} = 8.845$

$$t_{1,3} = \frac{\bar{y}_{1.} - \bar{y}_{3.}}{\sqrt{MSR}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{6.08 - 5.84}{0.2324\sqrt{\frac{1}{5} + \frac{1}{5}}} = 1.633$$

$$t_{2,3} = \frac{\bar{y}_{2.} - \bar{y}_{3.}}{\sqrt{MSR}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{4.78 - 5.84}{0.2324\sqrt{\frac{1}{5} + \frac{1}{5}}} = -7.212$$

$$\begin{aligned}
SSR &= \sum_{i=1}^{3}(n_i - 1)s_i^2 = 4(0.022 + 0.097 + 0.043) = 0.648 \\
SSR &\overset{or}{=} SST_o - SST = 5.4333 - 4.7853 = 0.648 \\
MSR &= \frac{SSR}{N - g} = \frac{0.648}{15 - 3} = 0.054 = 0.2324^2
\end{aligned}$$

where

$$\begin{aligned}
SST &= \sum_{i=1}^{3} n_i(\bar{y}_{i.})^2 - N\bar{y}^2 = 5(6.06^2 + 4.78^2 + 5.84^2) - 15(5.57)^2 = 4.785333 \\
SST_o &= \sum_{i=1}^{3}\sum_{j=1}^{n_i} y_{ij}^2 - N\bar{y}^2 = 470.25 - 15(5.57)^2 = 5.433333
\end{aligned}$$

3. **Assumption:** $Y_{ij} \sim \mathcal{N}(\mu_i, \sigma^2)$ and Y_{ij} are independent.
4. P-value: p-value $= 2\,Pr(t_{12} \geq 8.845) < 0.002$.

$$p\text{-value} = 2\,Pr(t_{12} \geq 1.633) > 0.1.$$

$$p\text{-value} = 2\,Pr(t_{12} \leq -7.212) < 0.002.$$

   (d.f. $= N - g = 15 - 3 = 12,\ t_{12,0.001} = 3.93,\ t_{12,0.05} = 1.782$)
5. **Decision:** The level of significant for each pair of test is $\alpha^* = 0.05/3 = 0.017$. The mean quantity of dissolved oxygen between Location 1 and 2 and between Location 2 and 3 are different.

2. (a) To minimize $var(\bar{X} - \bar{Y})$, we have to choose

$$n_1 = n\,\frac{\sigma_x}{\sigma_x + \sigma_y} = 90 \times \frac{3}{3+5} = 33.75 \quad \text{or} \qquad 34,$$

and $n_2 = 90 - 34 = 56$.

(b) If $n_1 = 34$ and $n_2 = 56$, then

$$var(\bar{X} - \bar{Y}) = \frac{\sigma_x^2}{n_1} + \frac{\sigma_y^2}{n_2} = \frac{9}{34} + \frac{25}{56} = 0.711.$$

In order to achieve this same bound with $n_1 = n_2 = m$, we must have

$$\frac{\sigma_x^2}{m} + \frac{\sigma_y^2}{m} = 0.711 \Rightarrow \frac{9}{m} + \frac{25}{m} = 0.711$$

or $\dfrac{34}{m} = 0.7111$ or $m = 47.8$. We choose $n_1 = n_2 = 48$. Note that $n_1 + n_2 = 96 > 90$ in (a) because this is NOT an optimal allocation.

3. We want to show that in case of no ties, the Kruskal-Wallis test statistic can be written as

$$k_0 \quad = \quad \frac{12}{N(N+1)} \sum_{i=1}^{g} n_i\,\bar{r}_i^2 - 3(N+1).$$

using the given result

$$\sum_{i=1}^{g}\sum_{j=1}^{n_i} r_{ij}^2 = \sum_{i=1}^{N} i^2 = \frac{1}{6}N(N+1)(2N+1);$$

and some basic result,

$$\sum_{i=1}^{g}\sum_{j=1}^{n_i} r_{ij} = \sum_{i=1}^{N} i = N(N+1)/2, \qquad \bar{r} = (N+1)/2$$

This implies that

$$
\begin{aligned}
\sum_{i=1}^{g} n_i(\bar{r}_{i\cdot} - \bar{r})^2 &= \sum_{i=1}^{g} n_i(\bar{r}_{i\cdot})^2 - N(\bar{r})^2 \\
&= \sum_{i=1}^{g} n_i(\bar{r}_{i\cdot})^2 - \frac{1}{4}N(N+1)^2, \\
\sum_{i=1}^{g}\sum_{j=1}^{n_i}(r_{ij} - \bar{r})^2 &= \sum_{i=1}^{g}\sum_{j=1}^{n_i} r_{ij}^2 - N(\bar{r})^2 \\
&= \frac{1}{6}N(N+1)(2N+1) - \frac{1}{4}N(N+1)^2 \\
&= \frac{1}{12}N(N+1)\left[2(2N+1) - 3(N+1)\right] \\
&= \frac{1}{12}N(N+1)(N-1).
\end{aligned}
$$

Now it follows easily that

$$
\begin{aligned}
k_0 &= (N-1)\frac{\sum_{i=1}^{g} n_i(\bar{r}_{i\cdot} - \bar{r})^2}{\sum_{i=1}^{g}\sum_{j=1}^{n_i}(r_{ij} - \bar{r})^2} \\
&= (N-1)\frac{\sum_{i=1}^{g} n_i(\bar{r}_{i\cdot})^2 - \frac{1}{4}N(N+1)^2}{\frac{1}{12}N(N+1)(N-1)} \\
&= \frac{12}{N(N+1)}\sum_{i=1}^{g} n_i\,(\bar{r}_{i\cdot})^2 - 3(N+1).
\end{aligned}
$$

4. Assuming $\underset{\sim}{g} = 2$, we want to show that the Kruskal-Wallis test statistic $K = \widetilde{W}^2$, where the $\widetilde{W}$ is the standardised Wilcoxon test statistic defined by

$$
\widetilde{W} = \frac{n_1\bar{r}_{1\cdot} - \frac{n_1(N+1)}{2}}{\sqrt{\frac{n_1 n_2(N+1)}{12}}}.
$$

where $n_1 + n_2 = N$, $\bar{r} = (n_1\bar{r}_{1\cdot} + n_2\bar{r}_{2\cdot})/N = (N+1)/2$.

When $g = 2$, the Wilcoxon test statistic $W = n_1\bar{r}_{1\cdot}$.

$$\begin{aligned}
E(r_{ij}) &= \frac{N(N+1)}{2} \cdot \frac{1}{N} = \frac{N+1}{2} \\
E(W) &= E(\sum_{j=1}^{n_1} r_{1j}) = n_1 E(r_{1j}) = \frac{n_1(N+1)}{2} \\
Var(r_{ij}) &= E(r_{ij}^2) - [E(r_{ij})]^2 = \frac{N(N+1)(2N+1)}{6} \cdot \frac{1}{N} - \left(\frac{N+1}{2}\right)^2 \\
&= \frac{(N+1)[2(2N+1) - 3(N+1)]}{12} = \frac{(N+1)(N-1)}{12} = \frac{N^2-1}{12} \\
Var(W) &= Var(\sum_{j=1}^{n_1} r_{1j}) = \sum_{j=1}^{n_1} Var(r_{1j}) + 2\sum_{j<j'}^{n_1} Cov(r_{1j}, r_{1j'}) \\
&= n_1 Var(r_{1j}) + n_1(n_1-1)Cov(r_{1j}, r_{1j'})
\end{aligned}$$

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

$$\begin{aligned}
W &= \sum_{i=1}^{N} i = \frac{1}{2}N(N+1) \quad \text{a constant} \\
\Rightarrow Var(W) &= NVar(r_{1j}) + N(N-1)Cov(r_{1j}, r_{1j'}) \\
\Rightarrow 0 &= NVar(r_{1j}) + N(N-1)Cov(r_{1j}, r_{1j'}) \\
\Rightarrow Cov(r_{1j}, r_{1j'}) &= \frac{-Var(r_{1j})}{N-1}
\end{aligned}$$

Hence we have

$$\begin{aligned}
Var(W) &= n_1 Var(r_{1j}) - \frac{n_1(n_1-1)}{N-1} Var(r_{1j}) \\
&= \frac{n_1(N-1) - n_1(n_1-1)}{N-1} Var(r_{1j}) \\
&= \frac{n_1(N-n_1)}{N-1} Var(r_{1j}) = \frac{n_1(N-n_1)}{N-1} \cdot \frac{N^2-1}{12} \\
&= \frac{n_1(N-n_1)}{N-1} \times \frac{(N+1)(N-1)}{12} = \frac{n_1 n_2(N+1)}{12}
\end{aligned}$$

Hence the standardised Wilcoxon test statistic is

$$\widetilde{W} = \frac{W - E(W)}{\sqrt{Var(W)}} = \frac{n_1\bar{r}_{1\cdot} - \frac{n_1(N+1)}{2}}{\sqrt{\frac{n_1 n_2(N+1)}{12}}}.$$

Then with no ties, we have

$$
\begin{aligned}
k_0 &= (N-1)\frac{\sum\limits_{i=1}^{g} n_i(\bar{r}_{i\cdot}-\bar{r})^2}{\sum\limits_{i=1}^{g}\sum\limits_{j=1}^{n_i}(r_{ij}-\bar{r})^2} \\
&= (N-1)\frac{n_1(\bar{r}_{1\cdot}-\bar{r})^2+n_2(\bar{r}_{2\cdot}-\bar{r})^2}{\frac{N(N+1)(N-1)}{12}} = \frac{n_1(\bar{r}_{1\cdot}-\bar{r})^2+n_2(\frac{N\bar{r}-n_1\bar{r}_{1\cdot}}{n_2}-\bar{r})^2}{\frac{N(N+1)}{12}} \\
&= \frac{n_1(\bar{r}_{1\cdot}-\bar{r})^2+\frac{1}{n_2}(N\bar{r}-n_1\bar{r}_{1\cdot}-n_2\bar{r})^2}{\frac{N(N+1)}{12}} = \frac{n_1(\bar{r}_{1\cdot}-\bar{r})^2+\frac{1}{n_2}(n_1\bar{r}-n_1\bar{r}_{1\cdot})^2}{\frac{N(N+1)}{12}} \\
&= \frac{n_1(\bar{r}_{1\cdot}-\bar{r})^2+\frac{n_1^2}{n_2}(\bar{r}-\bar{r}_{1\cdot})^2}{\frac{N(N+1)}{12}} = \frac{n_1n_2(\bar{r}_{1\cdot}-\bar{r})^2+n_1^2(\bar{r}-\bar{r}_{1\cdot})^2}{\frac{n_2N(N+1)}{12}} \\
&= \frac{n_1N(\bar{r}_{1\cdot}-\bar{r})^2}{\frac{n_2N(N+1)}{12}} = \frac{n_1(\bar{r}_{1\cdot}-\bar{r})^2}{\frac{n_2(N+1)}{12}} = \frac{(n_1\bar{r}_{1\cdot}-n_1\bar{r})^2}{\frac{n_1n_2(N+1)}{12}} \\
&= \left[\frac{W-E(W)}{\sqrt{Var(W)}}\right]^2 = \widetilde{W}^2
\end{aligned}
$$

Extra problems

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

Diet 1	Diet 2	Diet 3	Diet 4
3	9.5	5.5	18
1	2	12	7
8	4	17	14
15	12	9.5	5.5
	16		12

The KW test for the equality of the mean amounts of weight loss in pounds for the four diets is

1. **Hypotheses:** $H_0: \mu_1 = ... = \mu_g$ vs
   H_1 : Not all the μ_i's are equal.
2. **Test statistic:**

$$
\begin{array}{llllllll}
n_i & 4 & 5 & 4 & 5 & N = 18 & g = 4 \\
\bar{r}_{i.} & 6.75 & 8.7 & 11.0 & 11.3 & \bar{r} = 9.5 & \sum_{i=1}^{g}\sum_{j=1}^{n_i} r_{ij}^2 = 2106
\end{array}
$$

$$\sum_{i=1}^{g} n_i \bar{r}_i^2 - N\bar{r}^2 = 4(6.75^2) + 5(8.7^2) + 4(11^2) + 5(11.3^2) - 18(9.5^2) = 58.65$$

$$\sum_{i=1}^{g}\sum_{j=1}^{n_i} r_{ij}^2 - N\bar{r}^2 = 2106 - 18(9.5^2) = 481.5$$

$$k_0 = (N-1)\frac{\sum_{i=1}^{g} n_i \bar{r}_i^2 - N\bar{r}^2}{\sum_{i=1}^{g}\sum_{j=1}^{n_i} r_{ij}^2 - N\bar{r}^2} = 17\,\frac{58.65}{481.5} = 2.0707.$$

3. **Assumption:** No particular assumption on Y_{ij}. We have $k_0 \sim \chi^2_{g-1}$ under H_0.
4. ***P*-value:** p-value $= \Pr(\chi^2_3 \geq k_0) = \Pr(\chi^2_3 \geq 2.070717) > 0.01$ (0.5579 from R)
5. **Decision:** Since p-value > 0.05, we accept H_0. The data is consistent with H_0 that the mean amounts of weight loss in pounds for the four diets are equal.