THE UNIVERSITY OF SYDNEY

STAT2012 STATISTICAL TESTS

Semester 2 **Solution to Tutorial Week 5** 2015

Tutorial questions

1. The sample sizes, means and s.d. are

$$n_x = 13,\ n_y = 8,\ \bar{x} = 80.02,\ \bar{y} = 79.98,\ s_x = 0.02397 \text{ and } s_y = 0.03137.$$

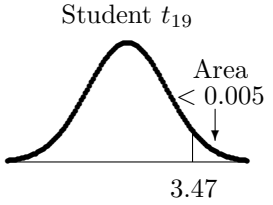
The *2 samples t-test* is

1. **Hypotheses:** $H_0 : \mu_x = \mu_y$ against $H_1 : \mu_x > \mu_y$.
2. **Test statistic:** $t_0 = \dfrac{\bar{x} - \bar{y}}{s_p\sqrt{\frac{1}{n_x} + \frac{1}{n_y}}} = \dfrac{80.02 - 79.98}{0.02693\sqrt{\frac{1}{13} + \frac{1}{8}}} = 3.4722.$

$$\begin{aligned} s_p^2 &= \frac{(n_x - 1)s_x^2 + (n_y - 1)s_y^2}{(n_x + n_y - 2)} \\ &= \frac{(13-1)0.02397^2 + (8-1)0.03137^2}{(13 + 8 - 2)} = 0.02693^2 \end{aligned}$$

3. **Assumption:** Assume that $X_i \sim \mathcal{N}(\mu_x, \sigma^2)$ & $Y_i \sim \mathcal{N}(\mu_y, \sigma^2)$. X_i and Y_i are independent. Then $t_0 \sim t_{n_x+n_y-2}$.
4. ***P*-value:** P-value $= P(t_{19} > 3.4722) \in (0.001, 0.005)$ (0.00127 from R).
5. **Decision:** Since the p-value is < 0.05, we reject H_0 and conclude that there is strong evidence against H_0 and conclude that Method A gives higher measurement.

Note: the p-value using WRS test in the lecture example is 0.00336 which is slightly higher because t test is more powerful in general.



2. We have $n_x = n_y = 5$ and $N = 10$. The ranks for the combined sample are

X	12	16	16	12	10	Y	30	12	24	32	24
Ranks	3.0	5.5	5.5	3.0	1.0	Ranks	9.0	3.0	7.5	10.0	7.5

The Wilcoxon rank-sum test for the difference between diet A and B is

1. **Hypotheses:** $H_0: \ \mu_x = \mu_y \quad \text{vs} \quad \mu_x \neq \mu_y$.
2. **Test statistic:** $W = 3.0 + 5.5 + 5.5 + 3.0 + 1.0 = 18$
3. **Assumption:** X_i and Y_i follow the same kind of distribution, differ by a shift.
4. ***P*-value:** From the WRS table, $\Pr(W \leq 18) = 0.0278$. However, this is only an approximation. We assume no ties when using the table. Since there are ties, we should use normal approximation to the distribution of W or derive the exact distribution of W. In this case, it doesn't matter whether one sum the ranks from the smaller or larger sample or whether the sum is in the lower or upper range.

$$\begin{aligned}
E(W) &= \frac{n_x(N+1)}{2} = \frac{5 \times (10+1)}{2} = 27.5 \\
g &= \frac{N(N+1)^2}{4} = \frac{10(10+1)^2}{4} = 302.5 \\
Var(W) &= \frac{n_x n_y}{N(N-1)} \left(\sum_{i=1}^{N} r_i^2 - \frac{N(N+1)^2}{4} \right) \\
&= \frac{5(5)(382 - 302.5)}{10(9)} \quad \text{since } \sum_{i=1}^{N} r_i^2 = 3^2 + \cdots + 7.5^2 = 382 \\
&= 22.083 \\
p\text{-value} &= 2\Pr(W \leq 18) = 2\Pr\left(Z < \frac{w - E(W)}{Var(W)} \right) \\
&= 2\Pr\left(Z < \frac{18 - 27.5}{\sqrt{22.083}} \right) = 2\Pr(Z < -2.021582) \\
&= 0.04321959
\end{aligned}$$

5. **Decision:** Since the p-value < 0.05. There is evidence in the data against H_0. There are differences in weights of pigs using diets X and Y.

Under no ties, we also have

$$\min = n_x \frac{1+N}{2} = \frac{5 \times (1+5)}{2} = 15, \quad \max = n_x \frac{n_y + 1 + N}{2} = \frac{5 \times (6+10)}{2} = 40$$

This is a rough sketch because, for example, min=1+3+3+3+5.5=15.5 and there are bars for non-integral rank sums.

Exact distribution in the presence of ties

We want the conditional distribution of W under H_0 given the observed set of average ranks:

$$1, \ 3, \ 3, \ 3, \ 5.5, \ 5.5, \ 7.5, \ 7.5, \ 9, \ 10.$$

Consider the ranks of X obtained by drawing 5 numbers from a box with numbers equal to ranks above. Thus starting from the lowest rank,

$$\begin{aligned}
\Pr(W = 15.5) &= \Pr(\text{ranks} = 1, 3, 3, 3, 5.5) = \frac{2}{\binom{10}{5}} \text{ (2 choices for 5.5)} \\
\Pr(W = 17.5) &= \Pr(\text{ranks} = 1, 3, 3, 3, 7.5) = \frac{2}{\binom{10}{5}} \text{ (2 choices for 7.5)} \\
\Pr(W = 18) &= \Pr(\text{ranks} = 1, 3, 3, 5.5, 5.5) = \frac{3}{\binom{10}{5}} \text{ (3 choices for 3)}
\end{aligned}$$

Thus

$$\Pr(W \leq 18) = \frac{2+3+2}{\binom{10}{5}} = 0.0278 \text{ (0.0278 from table; 0.0216 under normal approx.)}$$

and p-value is $2\Pr(W \leq 18) = 2 \times 0.0278 = 0.0556$.

3. We have $m = 3$, $n = 3$, $N = 6$ and $\left(\begin{array}{c} N \\ n \end{array} \right) = \left(\begin{array}{c} 6 \\ 3 \end{array} \right) = 20$ possible cases.

Rank of X	W	Rank of X	W
1,2,3	6	4,5,6	15
1,2,4	7	3,5,6	14
1,3,4	8	2,5,6	13
1,2,5	8	3,4,6	13
1,2,6	9	3,4,5	12
1,3,5	9	2,4,6	12
2,3,4	9	1,5,6	12
1,3,6	10	2,4,5	11
1,4,5	10	2,3,6	11
2,3,5	10	1,4,6	11

Under H_0, each occurs with probability

$$\binom{6}{3}^{-1} = \left(\frac{6!}{3!3!}\right)^{-1} = \left(\frac{6 \times 5 \times 4 \times 3 \times 2 \times 1}{3 \times 2 \times 1 \times 3 \times 2 \times 1}\right)^{-1} = \frac{1}{20}.$$

Thus

$$\begin{aligned}
\Pr(W \leq 6) &= \Pr(W \geq 15) = \frac{1}{20} = 0.05, \\
\Pr(W \leq 7) &= \Pr(W \geq 14) = \frac{2}{20} = 0.10, \\
\Pr(W \leq 8) &= \Pr(W \geq 13) = \frac{4}{20} = 0.20, \\
\Pr(W \leq 9) &= \Pr(W \geq 12) = \frac{7}{20} = 0.35, \\
\Pr(W \leq 10) &= \Pr(W \geq 11) = \frac{10}{20} = 0.50, \\
\Pr(W \leq 11) &= \Pr(W \geq 10) = \frac{13}{20} = 0.65
\end{aligned}$$

The values agree with the exact distribution that appears in the WRS Table.

4. Given a set of ranks $r_i,\ i = 1, \ldots, N$ with ties from the combined sample, an estimate of the variance of r_i is

$$s_r^2 = \frac{1}{N-1}\left(\sum_{i=1}^{N} r_i^2 - N\bar{r}^2\right) = \frac{1}{N-1}\left(\sum_{i=1}^{N} r_i^2 - \frac{N(1+N)^2}{4}\right)$$

since the average rank is $\bar{r} = \frac{1+N}{2}$. Moreover since the test statistic $W = \sum_{i=1}^{n_x} r_i$ is the sum of a subset of n_x ranks from totally N ranks of r_i, $var(W) \approx n_x s_r^2$. Hence

$$var(W) = \left(1 - \frac{n_x}{N}\right) n_x s_r^2 = \frac{n_y}{N} n_x s_r^2 = \frac{n_x n_y}{N(N-1)}\left(\sum_{i=1}^{N} r_i^2 - \frac{N(1+N)^2}{4}\right)$$

where $\left(1 - \frac{n_x}{N}\right)$ is a *finite population correction* factor for a sample of size n_x from a population of size N. You will learn it in Sample Survey of 3rd year.

Extra problems

1. We have $m = 7$, $\overline{x} = 6.34286.$, $s_x^2 = 1.40102^2$, $n = 8$, $\overline{y} = 8.0625$ and $s_y^2 = 1.10446^2$. The 2 samples t-test for the cholesterol level of patients in the treatment and control groups is

   1. **Hypothesis:** $H_0 : \ \mu_x = \mu_y \quad \text{vs} \quad H_1 : \ \mu_x < \mu_y$
   2. **Test statistic:** $t_0 = \dfrac{\overline{x} - \overline{y}}{s_p\sqrt{\frac{1}{n_x} + \frac{1}{n_y}}} = \dfrac{6.34286 - 8.0625}{1.2501\sqrt{\frac{1}{7} + \frac{1}{8}}} = -2.6579$

$$\begin{aligned} s_p^2 &= \frac{(n_x - 1)s_x^2 + (n_y - 1)s_y^2}{n_x + n_y - 2} \\ &= \frac{(7-1) \times 1.40102^2 + (8-1) \times 1.10446^2}{7 + 8 - 2} \\ &= 1.2501^2 \end{aligned}$$

   3. **Assumptions:** $X_i \sim \mathcal{N}(\mu_x, \sigma^2)$, $Y_i \sim \mathcal{N}(\mu_y, \sigma^2)$ and X_i & Y_i are independent.
   4. **P-value:** $0.005 \ < \ p\text{-value} = \Pr(t_{13} < -2.6579) < \ 0.01$ (0.00986, from R)
   5. **Decision:** Reject H_0 and conclude that the cholesterol levels of patients in the treatment group are lower than those in the control group.

2. The two samples t test without the assumption of equal variance is

   1. **Hypothesis:** H_0: $\mu_x - \mu_y = 0$ vs H_1: $\mu_x - \mu_y < 0$
   2. **Test statistic:** $t_0 = \dfrac{\overline{x} - \overline{y}}{\sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}} = \dfrac{6.34286 - 8.0625}{\sqrt{\frac{1.40102^2}{7} + \frac{1.10446^2}{8}}} = -2.6137$

$$\begin{aligned} df &= \frac{\left[s_x^2/n_x + s_y^2/n_y\right]^2}{\frac{(s_x^2/n_x)^2}{n_x} + \frac{(s_y^2/n_y)^2}{n_y}} \\ &= \frac{\left[1.40102^2/7 + 1.10446^2/8\right]^2}{\frac{(1.40102^2/7)^2}{7-1} + \frac{(1.10446^2/8)^2}{8-1}} = 11.40808 \end{aligned}$$

   3. **Assumption:** $X_i \sim \mathcal{N}(\mu_x, \sigma_x^2)$, $Y_i \sim \mathcal{N}(\mu_y, \sigma_y^2)$ and X_i & Y_i are independent.
   4. ***P*-value:** $\Pr(t_{11.40808} < -2.6137) = 0.01173573$ (from R)
   5. **Decision:** Since p-value < 0.05, there is strong evidence in the data against H_0. The cholesterol levels of patients in the treatment group are lower than those in the control group.

3. The ranks for the combined sample are

Trained	78	64	75	45	82	Control	110	70	53	51
Ranks	7	4	6	1	8	Ranks	9	5	3	2

We look at the ranks for the smaller sample. The Wilcoxon rank-sum test is

1. **Hypotheses:** $H_0: \ \mu_x = \mu_y \quad \text{vs} \quad \mu_x \neq \mu_y$.
2. **Test statistic:** $W_y = 9 + 5 + 3 + 2 = 19$
3. **Assumption:** X_i and Y_i follow the same kind of distribution differ by a shift.
4. ***P*-value:** $2\Pr(W \leq 19) = 2 \times 0.4524 = 0.9048$ (Table, $n = 4, 5, \ w = 19$) where $E(W) = \dfrac{n_y(N+1)}{2} = \dfrac{4 \times 10}{2} = 20$ and so $W = 19 < E(W)$ lies in the lower range.
5. **Decision:** Since the p-value > 0.05. The data is consistent with H_0. There are no differences in the number of trials required between the trained rats and the controls.

If one consider $W_x = 7 + 4 + 6 + 1 + 8 = 26$, p-value$=2\Pr(W \geq 26)$ can't be found from WRS table with $n_1 = 5, n_2 = 4$ but it can be found using
`2*(1-pwilcox(26-15-1,5,4))=0.9048` in `R` since

$$E(W_y) = \frac{n_y(1+N)}{2} = \frac{5(9+1)}{2} = 25 \quad \text{and} \quad \min(W_y) = \frac{n_y(n_y+1)}{2} = \frac{5(5+1)}{2} = 15$$

and the sum of prob. in the upper side from 26-15 is 1 minus the sum of prob. in the lower side from 26-15-1.