Assignment 1 Answer Keys

1. $H_0: \mu_B = \mu_A$ vs $H_1: \mu_B > \mu_A$.

a) Based on $H_0$ and randomization, the following sequences are equally likely,

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th>$\bar{y}_A$</th>
<th>$\bar{y}_B$</th>
<th>$\bar{y}_B - \bar{y}_A$</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>B</td>
<td>B</td>
<td>A</td>
<td>B</td>
<td>2.00</td>
<td>6.00</td>
<td>4.00</td>
</tr>
<tr>
<td>B</td>
<td>B</td>
<td>A</td>
<td>A</td>
<td>B</td>
<td>3.00</td>
<td>5.33</td>
<td>2.33</td>
</tr>
<tr>
<td>B</td>
<td>A</td>
<td>B</td>
<td>A</td>
<td>B</td>
<td>3.00</td>
<td>5.33</td>
<td>2.33</td>
</tr>
<tr>
<td>A</td>
<td>A</td>
<td>B</td>
<td>B</td>
<td>B</td>
<td>4.00</td>
<td>4.67</td>
<td>0.67</td>
</tr>
<tr>
<td>A</td>
<td>B</td>
<td>A</td>
<td>B</td>
<td>B</td>
<td>4.00</td>
<td>4.67</td>
<td>0.67</td>
</tr>
<tr>
<td>B</td>
<td>B</td>
<td>B</td>
<td>A</td>
<td>A</td>
<td>4.50</td>
<td>4.33</td>
<td>-0.17</td>
</tr>
<tr>
<td>B</td>
<td>A</td>
<td>A</td>
<td>B</td>
<td>A</td>
<td>5.00</td>
<td>4.00</td>
<td>-1.00</td>
</tr>
<tr>
<td>A</td>
<td>B</td>
<td>B</td>
<td>B</td>
<td>A</td>
<td>5.50</td>
<td>3.67</td>
<td>-1.83</td>
</tr>
<tr>
<td>B</td>
<td>A</td>
<td>B</td>
<td>B</td>
<td>A</td>
<td>6.50</td>
<td>3.00</td>
<td>-3.50</td>
</tr>
<tr>
<td>B</td>
<td>B</td>
<td>A</td>
<td>B</td>
<td>A</td>
<td>6.50</td>
<td>3.00</td>
<td>-3.50</td>
</tr>
</tbody>
</table>

The column $\bar{y}_B - \bar{y}_A$ gives the possible values of the sample mean difference each with probability .10.

b) The first sequence (bold-faced) was used in the experiment, and the observed sample mean difference is 4. Its $P$-value is

$$P(\bar{Y}_B - \bar{Y}_A \geq 4.00) = 1/10 = .1.$$  

c) Two-sample $t$ test:

$$s_A^2 = 2, s_B^2 = 3, s_{pool}^2 = \frac{(2-1)s_A^2 + (3-1)s_B^2}{2 + 3 - 2} = 2.67.$$  

The observed $t$ statistic is

$$t_0 = \frac{\bar{y}_B - \bar{y}_A}{s_{pool} \sqrt{1/n_A + 1/n_B}} = 2.68.$$  

The $p$-value is

$$P(t \geq t_0) = P(t \geq 2.68 \mid t(n_A + n_B - 2)) = .04$$

d) The validity of $t$ test depends on a set of sample and distributional assumptions. When the sample size is large, $t$ test is a good approximation of randomization test. Since the
sample sizes are relatively small in this experiment, the observed significance based on the randomization test is more realistic. Using $\alpha = 10\%$, we conclude that factor B has bigger effect.

2.

a) Let $\tau_A$, $\tau_B$, and $\eta_i$ be the effects of keyboard A, keyboard B and the learning effect. The learning effect will be present only when the same manuscript was typed in the second time. The first sequence (without randomization) is


in which $\eta_i$ is present when keyboard B is used. A statistical model is

$$y_{iA} = \mu + \tau_A + \epsilon_{iA},$$

$$y_{iB} = \mu + \tau_B + \eta_i + \epsilon_{iB},$$

where $i = 1, 2, \ldots, 6$. The sample means difference, $\bar{y}_B - \bar{y}_A$ is used to estimate $\tau_B - \tau_A$, and its expectation

$$E(\bar{y}_B - \bar{y}_A) = (\tau_B + \tau_l) - \tau_A.$$ 

Hence $\bar{y}_B - \bar{y}_A$ is a biased estimate. The bias is $\eta_l$. The second sequence (with randomization) is


The learning effect is associated with keyboard B four times, and keyboard A two times. If the sample mean difference $\bar{y}_B - \bar{y}_A$ is used to estimate $\tau_B - \tau_A$,

$$E(\bar{y}_B - \bar{y}_A) = (\tau_B + \frac{2}{3} \tau_l) - (\tau_A + \frac{1}{3} \tau_l)$$

$$= \tau_B + \frac{1}{3} \tau_l - \tau_A$$

2
Hence with randomization, though the estimate is still biased, the bias has been reduced to \( \frac{1}{3} \tau_i \).

To eliminate the learning effect entirely, some restriction need be imposed on randomization. In this example, balancing is required, that is to include equal number of \( A - B \)’s and \( B - A \)’s. In the following is one of the possible sequences,

\[
\]

Using the similar statistical model, it can be shown that

\[
E(\bar{y}_B - \bar{y}_A) = (\tau_B + \frac{1}{3} \tau_i) - (\tau_A + \frac{1}{3} \tau_i) = \tau_B - \tau_A
\]

b)

\[
\]

There exists an apparent pattern in the above sequence, that is, three \( A - B \)’s are followed by three \( B - A \)’s. This pattern may cause the estimate to be biased. As suggested, this can happen if the learning effect is not a constant in the experiment. We assume that it decreases across experimental runs. Let \( \tau_{i1}, \tau_{i2}, \tau_{i3}, \tau_{i4}, \tau_{i5}, \tau_{i6} \) be the learning effects associated with the manuscripts respectively, and

\[
\tau_{i1} > \tau_{i2} > \tau_{i3} > \tau_{i4} > \tau_{i5} > \tau_{i6}.
\]

Then,

\[
E(\bar{y}_B - \bar{y}_A) = (\tau_B + \frac{\tau_{i1} + \tau_{i2} + \tau_{i3}}{3}) - (\tau_A + \frac{\tau_{i4} + \tau_{i5} + \tau_{i6}}{3})
\]

Thus the bias is

\[
\frac{\tau_{i1} + \tau_{i2} + \tau_{i3}}{3} - \frac{\tau_{i4} + \tau_{i5} + \tau_{i6}}{3}
\]

which is bigger than 0.
3.

Use proc reg, model statement, and test statement to fit the full model and test the submodel with several insignificant variable deleted. Extra sum of squares for the deleted variables can be calculated by SAS, and a general $F$ is used to test whether these variables contribute to the regression significantly. Usual residual plots should also be generated for checking assumptions.