Week 6 Lecture: Multiple Comparison Tests (Chapter 11)

If we reject the null hypothesis, $H_0: \mu_1 = \mu_2 = \mu_3 = \ldots = \mu_k$, we usually want to know where inequalities exist among the different $k$ means. This idea is analogous to that of subdividing chi-square tables. Many methods exist to detect differences between individual means, but we will look at two:

1. Tukey test
2. Student-Newman-Keul (SNK) test
3. Dunnett’s test
4. Scheffe’s Multiple Contrast.

Multiple comparison tests have the same assumptions of ANOVA: normality and homogeneity of variance. Thought these tests are somewhat robust, nonparametric multiple comparison tests exist if the assumptions are seriously violated. All multiple comparison tests work best if sample sizes are equal.

**Tukey Test**

The Tukey test is probably the most “conservative” multiple comparison test. It tests the two-tailed null hypothesis, $H_0: \mu_a = \mu_b$, where $a$ & $b$ represent all possible combinations of the $k$ sample means. Here’s the procedure:

1. Compute the standard error:

   a. For equal $n$: $s_\pi = \sqrt{\frac{s_p^2}{n}} = \sqrt{\frac{\text{MSE}}{n}}$

   b. For unequal $n$: $s_\pi = \sqrt{\frac{s_p^2}{2 \left( \frac{1}{n_a} + \frac{1}{n_b} \right)}}$
2. Rank the sample means from lowest to highest.

3. Compare \( |\bar{X}_a - \bar{X}_b| \) to \( q_{a,v,k} * s_\bar{x} \) and reject equality of means if \( |\bar{X}_a - \bar{X}_b| \geq q_{a,v,k} * s_\bar{x} \)

An equivalent expression is to reject \( H_0: \mu_a = \mu_b \) if

\[
|\bar{X}_a - \bar{X}_b| \geq q_{a,v,k}
\]

(see Table B.5 for critical q-values). Note that the proper procedure is to compare the largest mean against the smallest, & then proceed to compare the next to the smallest mean to the smallest. If any non-significant differences are detected between two means, then no comparisons are made for other means within this interval. (see Zar page 230, and examples 11.1 & 11.2).

**Example**: \( X_i \) = transformed counts per 400 cm\(^2\) of oysters in 6 locations on an artificial reef.

\( \alpha = 0.05, k = 6, n = 12, \text{MSE} = 10.7, \text{df} = N - k = 72 - 6 = 66. \)

Ranked means: 4.05, 7.76, 7.85, 10.48, 10.54, 11.28.

\[
s_\bar{x} = \sqrt{\frac{\text{MSE}}{n}} = \sqrt{\frac{10.7}{12}} = 0.944 \quad q_{a,v,k} = q_{0.05,66,6} = 4.163
\]

So, declare means significantly different if

\[
|\bar{X}_a - \bar{X}_b| \geq (4.163)(0.944) = 3.93 \text{ units.}
\]

Thus, any means closer together than 3.93 units cannot be statistically separated. Our final result is:

| 4.05 | 7.76 | 7.85 | 10.48 | 10.54 | 11.28 |

where the underlined means are those that are not significantly different. The double underline for means 7.76 and 7.85 means that a Type II error occurred. We can write our result as:

\( \mu_1 \neq \mu_4 \neq \mu_5 = \mu_6 \). Note that we can say nothing about \( m_2 \) and \( m_3 \), because of the Type II error.

We could repeat the experiment with more samples and possibly get more conclusive answers since increasing \( n \) increases power (and thus less chance of committing a Type II error).
If we had unequal sample sizes in this example, we would use the exact same procedure except that we would calculate the standard error with the other formula (see example 11.2).

**Student-Newman-Kuels (SNK) Test**

This test uses the same procedure as the Tukey test, except that a different q-value is used depending on the number of means in the range being tested (call this number “p”). So, as p changes, \( q_{\alpha,v,k} \) changes depending on which means are compared.

**Example**: Back to our previous example…let’s redo using the SNK procedure.

- \( \mu_6 - \mu_1 \): 11.28 – 4.05 = 7.23. Compare to \( q_{0.05,6,66} \) * \( s_{\bar{x}} \) = 4.163 * 0.944 = 3.93. Since 7.23 > 3.93, we conclude that \( \mu_6 \neq \mu_1 \).

- \( \mu_6 - \mu_2 \) and \( \mu_5 - \mu_1 \):
  - For \( \mu_6 - \mu_2 \), 11.28 – 7.76 = 3.52. Compare to \( q_{0.05,6,66} \) * \( s_{\bar{x}} \) = 3.977 * 0.944 = 3.75. Since 3.52 < 3.75, we conclude that \( \mu_6 = \mu_2 \).
  - For \( \mu_5 - \mu_1 \), 10.54 – 4.05 = 6.49 > 3.75, so \( \mu_5 \neq \mu_1 \).
  - Since we did not detect any significant difference between the largest mean (6) and the second smallest mean (2), we will not make any further comparisons of means within that range.

- \( \mu_4 - \mu_1 \): 10.48 – 4.05 = 6.43. Compare to \( q_{0.05,6,66} \) * \( s_{\bar{x}} \) = 3.737 * 0.944 = 3.53. Since 6.43 > 3.53, we conclude that \( \mu_4 \neq \mu_1 \). Note that we do not compare \( \mu_5 - \mu_2 \) and \( \mu_6 - \mu_3 \).
• $\mu_3 - \mu_1: 7.85 - 4.05 = 3.80$. Compare to $q_{0.05,66,3} * s_\bar{\epsilon} = 3.399 * 0.944 = 3.21$. Since 3.80 $> 3.21$, we conclude that $\mu_3 \neq \mu_1$. Note that we do not compare $\mu_4 - \mu_2$, $\mu_5 - \mu_3$, and $\mu_6 - \mu_4$.

• $\mu_2 - \mu_1: 7.76 - 4.05 = 3.71$. Compare to $q_{0.05,66,2} * s_\bar{\epsilon} = 2.829 * 0.944 = 2.67$. Since 3.71 $> 2.67$, we conclude that $\mu_2 \neq \mu_1$. Note that we do not compare $\mu_3 - \mu_2$, $\mu_4 - \mu_3$, $\mu_5 - \mu_4$, and $\mu_6 - \mu_5$.

Our final results of the SNK test are:

<table>
<thead>
<tr>
<th></th>
<th>4.05</th>
<th>7.76</th>
<th>7.85</th>
<th>10.48</th>
<th>10.54</th>
<th>11.28</th>
</tr>
</thead>
</table>

The SNK test tends to reject more often than the Tukey test because it is more powerful. Some statisticians do not recommend using the SNK test because it may falsely detect significant differences with a probability greater than $\alpha$. Zar recommends the Tukey test over the SNK test because of this reason.

**Confidence Intervals After a Multiple Comparison Test**

We can compute a 95% confidence interval for two or more means that we conclude are the same. We first calculate a pooled mean:

$$
\bar{X}_p = \frac{\sum n_i \cdot \bar{X}_i}{\sum n_i}.
$$

With this pooled mean, we can then find the 95% confidence interval:

$$
\bar{X}_p \pm t_{\alpha(2),v} \sqrt{\frac{s_p^2}{\sum n_i}}.
$$
**Example:** Using our previous example, let’s find the confidence interval for $\mu_{2,3,4,5,6}$. First, get the pooled mean:

$$\bar{X}_p = \frac{(12 \times 7.76) + (12 \times 7.85) + (12 \times 10.48) + (12 \times 10.54) + (12 \times 11.48)}{(12 + 12 + 12 + 12 + 12)} = 9.58$$

Then, the 95% confidence interval is:

$$9.58 \pm t_{0.05(2),60} \sqrt{\frac{10.7}{60}} \Rightarrow 9.58 \pm (1.997)(0.4223) \Rightarrow 9.58 \pm 0.84.$$ 

We can also compute 95% confidence intervals for difference between means using the Tukey procedure: $$(\bar{X}_b - \bar{X}_a) \pm q_{a,v,k} \times s_{\bar{X}}.$$ Note that the standard error will be calculated differently depending on whether or not the sample sizes are equal (see previous discussion of Tukey test procedure).

**Example:** Continuing with our previous example, let’s find the confidence interval for $\mu_{2,3,4,5,6} - \mu_1$:

$$(9.58 - 4.05) \pm (4.163) \sqrt{\frac{10.7}{2} \left( \frac{1}{60} + \frac{1}{12} \right)} \Rightarrow 5.53 \pm 3.04$$

You should also see Zar’s example 11.3 for more examples.

**Dunnett’s Test**

This test is used to determine if several sample means are different than a control. Zar outlines the procedure in section 11.3 (page 234) with an example on page 236.
Scheffe’s Test

This test can be used like Tukey’s and SNK (though it is inclined to commit Type II errors), but it usually finds use as a “multiple contrast” test. A contrast is a linear function of the means \( \mu_i \); think of it as a difference between two means. Thus, Scheffe’s test can be used to determine if a group of means are significantly different from another group of means. For example, let’s say you have 3 means: \( \mu_1, \mu_2, \mu_3 \). Then, \( \text{Ho: } \frac{\mu_1 + \mu_2}{2} = \mu_3 \) is a contrast because it can be written as \( \text{Ho: } \frac{1}{2}\mu_1 + \frac{1}{2}\mu_2 + (-1)\mu_3 = 0 \). Scheffe’s test is the only test that allows you to compare groups of means with other groups of means.

The procedure to test a contrast is:

1. Set up the null hypothesis: \( \text{Ho: } c_1\mu_1 + c_2\mu_2 + \ldots + c_k\mu_k = 0 \)

2. Calculate the standard error as: \( s_\bar{x} = \sqrt{s_p^2 \left( \frac{c_1^2}{n_1} + \frac{c_2^2}{n_2} + \ldots + \frac{c_k^2}{n_k} \right)} \)

3. Calculate the test statistic, \( S \), as: \( S = \frac{c_1\bar{X}_1 + c_2\bar{X}_2 + \ldots + c_k\bar{X}_k}{s_\bar{x}} \)

4. Compare \( S \) to \( S_\alpha = \sqrt{(k-1)F_{\alpha}(1)(N-k)} \)

Note: This test does not require equal \( n_i \).

Example: Continuing with our previous example…let’s compare (\( \mu_2, \mu_3 \)) to (\( \mu_4, \mu_5, \mu_6 \)). We will test the null hypothesis:

\( \text{Ho: } 0\mu_1 + \frac{1}{2}\mu_2 + \frac{1}{2}\mu_3 + -1/3\mu_4 -1/3\mu_5 -1/3\mu_6 = 0, \)

or alternatively,
\[
\text{Ho: } \frac{\mu_2 + \mu_3}{2} = \frac{\mu_4 + \mu_5 + \mu_6}{3}.
\]

The standard error is:

\[
s_x = \sqrt{10.7 \left( \frac{1/4 + 1/4 + 1/9 + 1/9}{12} \right)} = 0.8620063
\]

The test statistic is:

\[
S = \frac{7.76 + 7.85 - 10.48 + 10.54 + 11.28}{0.862} = 3.4338496
\]

The critical value is:

\[
S_{\alpha} = \sqrt{(6-1)F_{0.05(1),5,66}} = \sqrt{5 * 2.36} = 3.4351
\]

Since \(3.433849 < 3.4351\) (just barely!), we do not reject Ho. This is a borderline case!!!

**Non-parametric Multiple Comparison Tests**

We can also perform multiple comparison tests for the Kruskal-Wallis test in a manner analogous to the Tukey and SNK tests. The non-parametric Tukey analog follows the same procedure as Tukey’s except for:

1. Use rank sums rather than sample means

2. Calculate the standard error as: \(s_x = \sqrt{\frac{n(nk)(nk+1)}{12}}\)

3. Compare to the critical value (from Table B.5): \(q_{\alpha,\nu,k}\)
Example: Here’s some hypothetical ranked data from \( k = 4 \) groups (n=4):

<table>
<thead>
<tr>
<th>Group A</th>
<th>Group B</th>
<th>Group C</th>
<th>Group D</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.5</td>
<td>9</td>
<td>11</td>
<td>12</td>
</tr>
<tr>
<td>5</td>
<td>10</td>
<td>8</td>
<td>15</td>
</tr>
<tr>
<td>3</td>
<td>1.5</td>
<td>7</td>
<td>14</td>
</tr>
<tr>
<td>4</td>
<td>6</td>
<td>16</td>
<td>13</td>
</tr>
</tbody>
</table>

\[ R_1 = 13.5 \quad R_2 = 26.5 \quad R_3 = 42 \quad R_4 = 54 \]

\[ s_\pi = \sqrt{\frac{n(nk)(nk + 1)}{12}} = \sqrt{\frac{4(4 * 4)(16 + 1)}{12}} = 9.522 \]

\[ q_{\alpha,\pi,k} = q_{0.05,\pi,4} = 3.633 \]

Decision Rule: If \( |R_a - R_b| \geq q_{\alpha,\pi,k} * s_\pi = (3.633)(9.522) = 34.59 \) then conclude that the two rank sums are significantly different.

Conclusion: From this decision rule, our final result is:

\[ 13.5 \quad 26.5 \quad 42 \quad 54 \]

We can also create an SNK non-parametric analog where the q-values change depending on the number of rank sums in the range being tested. In this case, we still multiply \( q * s_\pi \), but now:

1. \( q = q_{\alpha,\pi,p} \), where \( p = \) the number of rank sums in the range

2. \( s_\pi = \sqrt{\frac{n(np)(np + 1)}{12}} \).

Example: Continuing with our non-parametric example…

<table>
<thead>
<tr>
<th>P</th>
<th>( q = q_{0.05,\pi,p} )</th>
<th>( s_\pi )</th>
<th>( q * s_\pi )</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>3.633</td>
<td>9.522</td>
<td>34.59</td>
</tr>
<tr>
<td>3</td>
<td>3.314</td>
<td>7.211</td>
<td>23.90</td>
</tr>
<tr>
<td>2</td>
<td>2.772</td>
<td>4.899</td>
<td>13.58</td>
</tr>
</tbody>
</table>
- **R4 – R1**: 54 – 13.5 = 40.5 > 34.59 ***
- **R4 – R2**: 54 – 26.5 = 27.5 > 23.90 ***
- **R3 – R1**: 42 – 13.5 = 28.5 > 23.90 ***
- **R4 – R3**: 54 – 42 = 12.0 < 13.58 ns
- **R3 – R2**: 42 – 26.5 = 15.5 > 13.58 ***
- **R2 – R1**: 26.5 – 13.5 = 13.0 < 13.58 ns

So, our final result is (notice that there’s more separation than with the non-parametric Tukey procedure:

| 13.5 | 26.5 | 42  | 54  |

When we have **unequal sample sizes**, we can no longer use ranks sums. Rather, we must use rank means, since they are adjusted for sample size. This procedure is called **Dunn’s Test**, and it uses the test statistic:

\[
Q = \frac{|R_b - R_a|}{s_x},
\]

where

\[
s_x = \sqrt{\frac{N(N + 1)}{12} \left( \frac{1}{n_a} + \frac{1}{n_b} \right)}.
\]

We then compare the Q statistic to the critical value \(Q_{a,k}\) (see Table B.15). If tied ranks are present, then use the following equation to calculate the standard error (see page 241):

\[
s_x = \sqrt{\frac{N(N + 1)}{12} - \frac{\sum_{t} t}{12(N - 1)} \left( \frac{1}{n_a} + \frac{1}{n_b} \right)},
\]
where t is used for ties in the KW test (see equation 10.42 in Zar). Zar presents an example of Dunn’s test on page 242 (Example 11.8).

Zar also presents a non-parametric analog of Dunnett’s Test for comparing a control to other treatment groups (page 243) and a non-parametric analog for testing multiple contrasts (page 243), neither of which I will cover in lecture.

Also, SAS recommends using the REGWQ & REGWF multiple comparison tests (not mentioned in Zar) when you do not need confidence intervals and the sample sizes are equal. When you have unequal sample sizes or need confidence intervals, SAS recommends the Tukey test. They further recommend some additional tests (i.e., Waller-Duncan) based on a Bayesian approach that is not covered in Zar. They do not favorably recommend Scheffe’s multiple contrast or Duncan’s multiple range tests (the latter is briefly mentioned in Zar on page 232).

As you can see, there is a great variety of choice on which multiple comparison test to use. You may want to use two or three different tests when doing your projects just to get some different “viewpoints”.