10.3 WHICH MEANS ARE DIFFERENT?: MULTIPLE COMPARISONS

When an analysis of variance $F$-test indicates a significant difference among population means, then the experimenter naturally wants to determine which means are different. An investigation as to which means differ requires us to compare all pairs of population means. For example, consider the simplest situation in which there are three population means, and the hypotheses for an analysis of variance are

$$H_0: \mu_1 = \mu_2 = \mu_3$$

$$H_a: \text{At least two population means differ.}$$

If we have concluded that the alternative is true, and want to know which means differ, we need to compare the following pairs of means: $\mu_1$ to $\mu_2$, $\mu_1$ to $\mu_3$, and $\mu_2$ to $\mu_3$. Since we want to determine whether there is sufficient evidence to conclude that there is a difference in any of these pairs of means, we need to test each of the following pairs of hypotheses.

**Pair 1:** $H_0: \mu_1 - \mu_2 = 0 \quad H_a: \mu_1 - \mu_2 \neq 0$

**Pair 2:** $H_0: \mu_1 - \mu_3 = 0 \quad H_a: \mu_1 - \mu_3 \neq 0$

**Pair 3:** $H_0: \mu_2 - \mu_3 = 0 \quad H_a: \mu_2 - \mu_3 \neq 0$

Since multiple tests are required to compare several pairs of means, any procedure that compares all these pairs is called a **multiple comparison** procedure. There are a number of different multiple comparison procedures that allow us to test the hypotheses of interest. We will use one of the most basic multiple comparison procedures, one that is a modification of the two-sample $t$-statistic that we used to compare two means (see Section 9.2).

Recall from Section 9.2 that the $t$-statistic is

$$t = \frac{(\overline{x}_1 - \overline{x}_2) - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}.$$  

When performing an analysis of variance, the "$s^2$" that measures the amount of variability in the samples is the expression appearing in the denominator of the $F$-statistic. As we noted at the end of Section 10.2, this quantity is called the **mean square error**, or MSE. It follows that the $t$-statistic that we will use to compare pairs of population means for a completely randomized design is given by

$$t = \frac{(\overline{x}_1 - \overline{x}_2) - 0}{\sqrt{\frac{MSE}{n_1} + \frac{MSE}{n_2}}}.$$
In Section 9.2 we used a form of the two-sample t-test that did not assume that the population variances are equal. However, one of the conditions required to validly conduct an analysis of variance is that the population variances are equal. We will conduct the multiple comparison t-tests under this condition that the population variances are equal. In this case, the test statistic of interest has a t-distribution with the number of degrees of freedom associated with the mean square error, MSE. In general these degrees of freedom may be expressed as $N - k$, where $N$ is the total number of measurements sampled, and $k$ is the number of means, or populations, being compared.

The general calculations required for conducting multiple comparison t-tests are summarized in the box below.

**Calculations for Multiple Comparisons of $k$ Population Means**

For each pair of means you wish to compare, calculate the following test statistic, and the corresponding p-value.

$$ t = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{\frac{MSE}{n_1} + \frac{MSE}{n_2}}} $$

The test statistic and $p$-value may be obtained with the aid of technology, or the $p$-value is obtained from Table 4 using $N - k$ degrees of freedom.

Because of the many comparisons to be made, we will relax the usual structure for testing hypotheses in this section. In particular we will not state the multiple hypotheses to be tested. It is understood that we will test all pairs of population means for possible differences. Also, we will not explicitly state either a test statistic or a decision rule. We will always use the test statistic appearing above to perform the required calculations. We will discuss how to interpret these results in later examples.

The conditions required for the validity of this process are identical to those required for an analysis of variance F-test. For emphasis, we reiterate these conditions are repeated in the box below.
Validity Conditions for the Multiple Comparison of $k$ Population Means

To validly compare all pairs of population means, the following conditions must be met.

1. A completely randomized design is used to collect the $k$ samples of data.
2. All $k$ populations are normally distributed.
3. All $k$ populations have the same standard deviation. That is, $\sigma_1 = \sigma_2 = \ldots = \sigma_k$.
4. The values of the population standard deviations are not known.

**EXAMPLE 10.4**

In Example 10.2 we concluded, at the .05 level of significance, that there are differences among the mean grade point averages of first-year college students in four socioeconomic groups. Which group has the lowest average GPA? Perform each analysis at the .05 level of significance. For convenience, these data are repeated in data file: exam10-04.

**Solution**

Since the conditions for conducting a valid analysis of variance were satisfied (see Example 10.2 in the preceding section), we may use the multiple comparison procedure described in the last box. We used technology, Minitab again, to calculate all the values of the test statistic and p-values. The results of these calculations are summarized in the second and third columns following table. The interpretations shown in the fourth column will be explained below.

<table>
<thead>
<tr>
<th>Population Means Compared</th>
<th>Value of $t$</th>
<th>p-value</th>
<th>Interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lower Middle vs. Poor</td>
<td>+1.41</td>
<td>0.167</td>
<td>NS</td>
</tr>
<tr>
<td>Lower Middle vs. Upper Middle</td>
<td>-1.05</td>
<td>0.301</td>
<td>NS</td>
</tr>
<tr>
<td>Lower Middle vs. Well-to-do</td>
<td>-1.26</td>
<td>0.216</td>
<td>NS</td>
</tr>
<tr>
<td>Poor vs. Upper Middle</td>
<td>-2.46</td>
<td>0.019</td>
<td>Poor $&lt;$ Upper Middle</td>
</tr>
<tr>
<td>Poor vs. Well-to-do</td>
<td>-2.67</td>
<td>0.011</td>
<td>Poor $&lt;$ Well-to-do</td>
</tr>
<tr>
<td>Upper Middle vs. Well-to-do</td>
<td>+0.21</td>
<td>0.835</td>
<td>NS</td>
</tr>
</tbody>
</table>

Let us carefully interpret these analyses. The first two population means being compared are the population mean GPA for first-year students in the Lower Middle class to those in the Poor class. The calculated value of the test statistic for this comparison is positive. This is because the sample mean for students in the Lower Middle Class is larger than the sample mean of those in the Poor class. Since this the p-value of 0.167 > $\alpha = 0.05$, there is insufficient evidence to conclude these population means are different. On the other hand, when comparing the population mean GPA of first-year students in the Poor class to those in the Upper Middle class,
the t-value is negative, and the p-value of $0.019 < \alpha = 0.05$. The t-value is negative because the sample mean for students in the Poor Class is smaller than the sample mean of those in the Upper Middle class. Based on the relative values of the sample means, and the fact that the p-value is less than the value of $\alpha$, there is sufficient evidence to conclude that the mean GPA for the Poor class is less than the mean GPA of the Upper Middle class. The other calculated tests and p-values have similar interpretations similar to the two discussed. The preceding table completely summarizes all of the analyses.

The notation used in the “Interpretation” column of the table deserves further explanation. The notation “NS” stands for non-significant. It is shorthand to indicate that there is insufficient evidence to conclude that there is any difference between the two population means being compared. “Poor<Upper Middle” denotes that the mean GPA for the poor class is less than that for the Upper Middle class. Similarly, “Poor<Well-to-do” denotes that the mean GPA for the poor class is less than that for the Well-to-do class. Remember that these statements are inferences about the respective population means.

Now, how do we use all of this information to answer the question as to which group has the lowest average GPA? Since we have concluded that the Poor class has a lower mean GPA than the Upper Middle class, we must conclude that the Upper Middle class cannot have the lowest mean GPA. Similarly, since we have concluded that the Poor class has a lower mean GPA than the Well-to-do class, we must conclude that the Well-to-do class cannot have the lowest mean GPA. Thus we have eliminated two of the groups from having the lowest mean GPA. Since mean GPAs for the other two classes are not significantly different, it may be the case that all students classified as Poor, or all students classified as Lower Middle class, have the lowest mean GPA. From this study it is not possible to determine which of these two might be lower.

The previous example demonstrates the basic reasoning we use with multiple comparisons. If we have concluded that there are differences in more than two treatments, it is generally of interest to determine which treatment is “best” or “worst.” That is, we would like to know which population mean might be the largest or smallest, depending on the meanings of “best” and “worst.” To do this, we must compare all possible pairs of population means. Based upon these comparisons, if we conclude that mean “A” is smaller than some other mean, then we immediately infer that “A” cannot be the largest mean. Similarly, if we conclude that mean “B” is larger than some other mean, then mean “B” cannot possibly be the smallest mean. If we are trying to determine the smallest mean, we will then eliminate any mean that is significantly larger than some other mean. Any remaining mean that we have not ruled out in this way could be the smallest mean. In an analogous fashion, if we are trying to determine the largest mean, then we eliminate any mean that is significantly smaller than some other mean. Any of the remaining means could be the largest.

We must clarify one final point. If the analysis of variance $F$-test indicates that there is insufficient evidence to conclude there are differences among the population means, then we will not perform a multiple comparison procedure. After all, it makes no sense to “see which means are different” if we are not convinced that some differences actually exist. In Example 10.3 we considered an experiment comparing five growth hormones for plants. In that case, there was insufficient evidence of any differences among the five growth hormones with respect to the mean weight increase of the shrub. Therefore, there is no reason to use multiple comparisons to determine which growth hormone is “best” or “worse.”
We now consider an example that ties together all these ideas.

**EXAMPLE 10.5**

The strength and endurance of women athletes is much studied. An experiment is conducted to compare the grip strengths of college women participating in five different sports: basketball (BB), Field Hockey (FH), Golf (G), Swimming (S), and Tennis (T). Physiologists randomly sample women athletes in each sport, and measure the maximal grip strength (in kilograms) for each. Data file: exam10-05. Conduct each of the following analyses at a .05 level of significance.

A. Is there sufficient evidence to conclude that there are some differences among the mean maximal grip strengths of women in the five sports?

B. If appropriate, determine which sport has women with the strongest grip.

C. If appropriate, determine which sport has women with the weakest grip.

**Solution**

A. Letting \( \mu_i \) represent the mean maximal grip strength of all basketball players, etc., the elements of the desired test are:

\[
H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 \\
H_a : \text{At least two of the population means are not equal}
\]

Decision Rule: Accept \( H_a \) if the p-value \( < .05 \)

Test Statistic: 
\[
F = \frac{\text{Variability among the sample means}}{\text{Variability expected by chance}}
\]

Using technology, we obtain the following analyses. The value of the \( F \)-statistic and the corresponding p-value are highlighted. Figure 10.3 shows a graphical representation of the value of the test statistic and the p-value for this example.
One-way ANOVA: Strength versus Sport

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sport</td>
<td>4</td>
<td>1237.5</td>
<td>309.4</td>
<td>9.04</td>
<td>0.000</td>
</tr>
<tr>
<td>Error</td>
<td>68</td>
<td>2326.2</td>
<td>34.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>72</td>
<td>3563.7</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

S = 5.849 R-Sq = 34.73% R-Sq(adj) = 30.89%

Individual 95% CIs For Mean Based on Pooled StDev

| Level | N   | Mean | StDev | +---------+---------+---------+---------|
|-------|-----|------|-------|----------|----------|----------|----------|
| BB    | 15  | 42.073| 5.384 | (-----*-----) |
| FH    | 18  | 41.217| 6.142 | (-----*-----) |
| G     | 13  | 42.569| 4.359 | (-----*------) |
| S     | 15  | 33.093| 7.195 | (-----*-----) |
| T     | 12  | 45.625| 5.445 | (-----*------) |

Pooled StDev = 5.849

Figure 10.3 Graphical Representation of the p-value for Example 10.5

Since the p-value of 0.000 is clearly less than .05, we accept the alternative hypothesis. At a .05 level of significance we may conclude that there are some differences among the mean maximal grip strengths for all women participating in the five sports.

B. Since we have concluded there are differences among the mean maximal grip strengths for the five sports, it is appropriate to use multiple comparisons to determine how the means differ. Again, Minitab was used to produce the analyses summarized in the following table.
### Population Means Compared

<table>
<thead>
<tr>
<th>Population Means Compared</th>
<th>Value of t</th>
<th>p-value</th>
<th>Interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>BB to FH</td>
<td>+0.42</td>
<td>0.676</td>
<td>NS</td>
</tr>
<tr>
<td>BB to G</td>
<td>-0.22</td>
<td>0.827</td>
<td>NS</td>
</tr>
<tr>
<td>BB to S</td>
<td>+4.20</td>
<td>0.000</td>
<td>BB &gt; S</td>
</tr>
<tr>
<td>BB to T</td>
<td>-1.57</td>
<td>0.121</td>
<td>NS</td>
</tr>
<tr>
<td>FH to G</td>
<td>-0.64</td>
<td>0.524</td>
<td>NS</td>
</tr>
<tr>
<td>FH to S</td>
<td>+3.97</td>
<td>0.000</td>
<td>FH &gt; S</td>
</tr>
<tr>
<td>FH to T</td>
<td>-2.02</td>
<td>0.047</td>
<td>FH &lt; T</td>
</tr>
<tr>
<td>G to S</td>
<td>+4.28</td>
<td>0.000</td>
<td>G &gt; S</td>
</tr>
<tr>
<td>G to T</td>
<td>-1.31</td>
<td>0.195</td>
<td>NS</td>
</tr>
<tr>
<td>S to T</td>
<td>-5.53</td>
<td>0.000</td>
<td>S &lt; T</td>
</tr>
</tbody>
</table>

The objective is to identify the sports with the largest mean grip strength. Thus, we want to eliminate those sports that have a mean that is significantly smaller than some other mean. Consulting the summary above, we see that we can eliminate swimming (BB > S, or FH > S, or G > S, or S < T), and field hockey (FH < T) by this reasoning. The means for the remaining sports are not significantly different. So we may infer that women participating in basketball, golf, or tennis could have the largest mean maximal grip strength.

C. To determine which sport has the smallest mean grip strength, we will eliminate those sports that have a mean that is significantly larger than some other mean. From the summary above, we can eliminate basketball (BB > S), field hockey (FH > S), golf (G > S), and tennis (S < T) from consideration. Since swimming is the only sport that we did not eliminate, we may infer that women swimmers have the smallest mean maximal grip strength.

We should conclude this discussion of multiple comparisons with a few final comments. The first of these has to do with the chance that a multiple comparison procedure will lead to an error. For convenience, assume that we used multiple t-tests, each with a level of significance of $\alpha = 0.05$, in our comparison procedure. For any one test of hypothesis, the probability of a Type I error, the probability of incorrectly concluding there is a difference between the two means we are comparing, is $\alpha = 0.05$. However, when multiple tests are performed, the probability of making at least one Type I error is not equal to $\alpha = 0.05$. In fact, the more tests that are done, the higher the risk is that at least one of them will be incorrect and falsely conclude that there is a difference in two population means, when there really is no difference. You may have noticed in the last two examples that we did not attach any statement of reliability to our final conclusions. The reason for this is that when we form multiple $t$-tests using data from a completely randomized design, there is no way to precisely calculate the chance that at least one of the tests has resulted in a Type I error.

The fact that we only use a multiple comparison procedure after a significant analysis of variance $F$-test helps to ensure that there really are some differences among the population means. Using multiple $t$-tests to compare multiple population means is a reasonable way to then sort out, in an orderly fashion, which population means might differ. As long as we are not comparing a very large number of means, and as long as we are careful to compare pairs of means only after a significant $F$-test, this procedure should not lead to many errors.
As a second comment, we observe that there exist other multiple comparison techniques that do control the “overall error rate.” The complexity of the theory behind these procedures is beyond the scope of this text. However, the reasoning used to interpret the results of any multiple comparison procedure should be exactly the same as reasoning discussed in this section. Readers interested in exploring other multiple comparison procedures should consult a more advanced text.¹

We conclude this section by summarizing when it is appropriate to use the multiple comparison method described above. An experimenter should consider the following items when deciding whether to use this multiple comparison procedure to determine which of \( k \) population means differ.

- The sampled data are obtained from the \( k \) populations using a completely randomized design.
- An analysis of variance \( F \) -test indicates that there are some differences among the \( k \) population means.
- The objective is to determine which of the \( k \) population means differ. It is usually of interest to determine which mean might be the largest (or smallest).

SECTION 10.3 EXERCISES

In each of the following exercises, assume that it is valid to conduct an analysis of variance \( F \) -test. You must be sure it is appropriate to perform multiple comparisons based upon the result of the analysis of variance \( F \) -test. You should be able to specify the conditions under which these statistical procedures are valid.

These exercises are continuations of the same numbered exercises from Section 10.2.

3. A research psychologist wants to investigate the difference in the lengths of times it takes a strain of laboratory mice trained under different laboratory conditions to reach the end of a maze. The psychologist randomly selects eighteen mice of this strain, with six receiving no training at all (control group), six trained under condition 1, and six trained under condition 2. She then places each mouse in the maze, and records the number of seconds it takes the mouse to reach the end. Data file: c10s03e03. [Control Group: \( \bar{x} = 50.333, s = 5.989 \); Condition 1: \( \bar{x} = 70.500, s = 7.791 \); Condition 2: \( \bar{x} = 68.667, s = 6.408 \) ] If appropriate, determine which laboratory condition is associated with the largest mean time to reach the end of the maze. Conduct all tests at a \( .05 \) level of significance.

6. In the introduction to this chapter we discussed a study of grades and work experience for a sample of 269 recent MBA graduates from the University of Louisville College of Business and Public Administration. For each student the MBA grade point average at graduation (GGPA), and the amount of work experience were recorded. The amount of work experience was categorized in one of three ways: one year or less, two or three years, or more than three years.² Data file: c10s03e06. [0 or 1

¹ For example, see Hsu, Jason C. (1996) Multiple Comparisons: Theory and methods, Chapman & Hall.
² College Student Journal, June 2000, Arthur J. Adams
Years Experience:  \( n = 79, \bar{x} = 3.45, s = 0.28 \); 2 or 3 Years Experience:  \( n = 77, \bar{x} = 3.55, s = 0.21 \); 4 or More Years Experience:  \( n = 113, \bar{x} = 3.64, s = 0.23 \)³

A. If appropriate, determine which category of work experience is associated with the highest mean GGPA. Conduct all tests at a .05 level of significance.

B. If appropriate, determine which category of work experience is associated with the lowest mean GGPA. Conduct all tests at a .05 level of significance.

CHAPTER 10 REVIEW EXERCISES

In each of the following exercises, assume that it is valid to conduct an analysis of variance \( F \)-test. You must be sure it is appropriate to perform multiple comparisons based upon the result of the analysis of variance \( F \)-test. You should be able to specify the conditions under which these statistical procedures are valid.

6. This exercise is a continuation of Exercise 3, Section 10.1. Psychologists have studied the effects of the working environment on the quality and quantity of work done. To improve the work environment, many businesses pipe music into work areas. To determine which type of music best enhances productivity, a certain company conducts the following experiment. Three types of music - country, rock, and classical - are played in the production area of the company, each on seven randomly selected days. Managers record the number of items produced at the end of each day. Data file: c10r06. [Country: \( \bar{x} = 799.86, s = 14.24 \); Rock: \( \bar{x} = 782.57, s = 12.04 \); Classical: \( \bar{x} = 859.71, s = 7.61 \)] Which music would you recommend this company use? Perform all appropriate analyses at the .05 level of significance.

8. Many employers have experimented with three different forms of the 40-hour workweek, in an effort to determine which form will maximize production and minimize expenses. A factory has tested a 5-day week (8 hours per day), a 4-day week (10 hours per day), and a 3\( \frac{1}{3} \)-day week (12 hours per day). Managers record the weekly production results (in thousands of dollars worth of items produced) for independent random samples of six weeks for each type of workweek. Data file: c10r08. [5-day week: \( \bar{x} = 102.67, s = 6.65 \); 4-day week: \( \bar{x} = 102.50, s = 8.67 \); 3\( \frac{1}{3} \)-day week: \( \bar{x} = 97.17, s = 4.58 \)] Which type of workweek would you recommend for this factory? Perform all appropriate analyses using a .05 level of significance.

³ The data in this example are not the actual data collected in the study, but have been constructed to reflect the same statistics as those summarized in the article.