# **Chapter 9**

# **Multiple Comparisons**

After performing an initial model test in an ANOVA situation, we often want to examine a set of treatments to determine where the differences lie. Suppose we have eight treatments to compare. If we compare all pairs of treatments, we have  = 28 contrasts to test. For 9 treatments we would have 36 tests, and for 10 treatments we would have 45 tests. It is very easy to accumulate a very large number of tests when examining data!

Suppose we perform each individual test at α = 0.05. What happens to our error rate with so many endpoint tests? It depends on how you define error rate. Suppose we define error rate to be the probability of at least one Type I error.



This is even worse for more treatments.

When we have a planned set of comparisons we wish to perform, we have two error rates:

1) the error rate for each individual comparison, and

2) the overall error rate for the entire set

Experimenters should determine which they wish to control based on the conclusions that will be reached.

Unless you use Scheffe’s method (discussed later), only pre-planned comparisons can be performed. (Comparisons inspired by the data are referred to as data-snooping. This will change your overall error rate.)

The number of comparisons to be made affects the overall error rate. Generally we are unsure what the overall error rate might be as it depends on the distribution of the sample means and the set of comparisons the researcher is interested in. There are, however, certain sets and situations where we can compute/control the overall error rate exactly. For other situations, we must consider conservative alternatives.

A conservative method is one in which we can be sure that the overall error rate is not greater than specified, say 0.05. Here the overall error rate could be considerably less than 0.05, but would never be greater than 0.05. We will consider some conservative methods and some exact methods.

Recall that there are three general procedures for performing multiple comparisons:

a) Simultaneous intervals—Tukey’s method for all pairwise where the means are independent, other methods for other sets. Allows estimation of the difference between the treatments. Usually easy to find conservative methods.

b) Range tests--Ryan-Einot-Gabriel-Welsch method for all pairwise, other methods for other sets. Only gives directional information, no estimates. Not as many options for conservative methods.

c) Ignore simultaneous procedures but make simultaneous conclusions anyway—LSD.

We will address them all in turn, but we will first consider c) for illustration purposes.

In a partial defense of option c), there are times when you are not concerned with the overall error rate, as you are not planning to pool information from your individual tests to draw overall conclusions. When this is the case, the method referred to as LSD (least significant difference) is appropriate. This is essentially one-at-a-time estimation or testing and technically cannot be referred to as multiple comparisons. LSD simply amounts to calculating confidence intervals based on the t-tests we discussed earlier for contrasts.

To begin with, we will suppose that our set of interest is all pairwise comparisons of the means. We will also let *s2* be an appropriate error term (often MSE) with an associated value for degrees of freedom.

Recall our t-statistic formula for contrasts: 

Here Q = , so we have .

We can “pull this equation apart” to obtain a confidence interval formula

 

for equal sample sizes this will become

.

We will consider two treatments to be significantly different from each other if the interval above does not contain zero. Thus if the sample sizes are equivalent we can compare the difference in the sample means to the quantity

.

If the absolute value of the difference is larger, the sample means are significantly different. Otherwise they are not significantly different.

Consider our first example: From the table we have, 

So we want to compare the mean differences to



Generally, we begin with the largest difference since if this difference is insignificant, all the other differences will be also. (Note that this procedure only works when the sample sizes are equal.)

B-D: 14.0929-10.91 = 3.1829, so B is significantly larger than D

B-A: 14.0929-11.9857 = 2.1072, so B is significantly larger than A

B-C: 14.0929-12.4286 = 1.6643, so B is not significantly different from C

C-D: 12.4286-10.91 = 1.5186, so C is not significantly different from D (and hence C and A are also not sig. different)

We can summarize using a line diagram:

D A C B

\_\_\_\_\_\_\_\_\_\_\_\_

Treatment B has a significantly higher average increase in haze than treatment D and treatment A. There are no other significant differences.

If we want to focus on the treatments with the smallest average increase in haze, we should eliminate treatment B from consideration, but no other treatments can be eliminated. (However, conclusions of this type are not valid here, as LSD does not control the overall error rate.)

We can also perform confidence intervals for the differences in the treatment means. For instance, for B-D we have



Now we will consider the second example. We will use 

In this instance, the quantity we need to compare the means to will change for each comparison. Thus, we will need to recompute the quantity each time.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Group: | 1 | 2 | 3 | 4 | 5 | 6 |
| Mean: | 4.05 | 7.76 | 7.85 | 10.48 | 10.54 | 11.28 |
| *ni*: | 9 | 15 | 12 | 16 | 8 | 12 |

We will use “\*” to indicate comparisons which are significant at the 0.05 level.

|(1)-(6)| = 7.23\* compare to 2.00 = 2.8844

|(1)-(5)| = 6.49\* compare to 2.00 = 3.1785

|(1)-(4)| = 6.43\* compare to 2.00 = 2.7255

|(1)-(3)| = 3.80\* compare to 2.00 = 2.8844

|(1)-(2)| = 3.71\* compare to 2.00 = 2.7580

|(2)-(6)| = 3.52\* compare to 2.00 = 2.5334

|(2)-(5)| = 2.78 compare to 2.00 = 2.8637

|(2)-(4)| = 2.72\* compare to 2.00 = 2.3509

|(2)-(3)| = 0.09 compare to 2.00 = 2.5334

|(3)-(6)| = 3.43\* compare to 2.00 = 2.6705

|(3)-(5)| = 2.69 compare to 2.00 = 2.9857

|(3)-(4)| = 2.63\* compare to 2.00 = 2.4980

|(4)-(6)| = 0.80 compare to 2.00 = 2.4980

|(4)-(5)| = 0.06 compare to 2.00 = 2.8324

|(5)-(6)| = 0.74 compare to 2.00 = 2.9857

Although not appropriate with this procedure, overall conclusions would be:

The mean density of oysters on a liberty ship artificial reef is lowest for the floors of holds. The mean density of oysters is significantly lower for the starboard deck than for the sides of holds or the port side. The mean density of oysters is significantly lower for the starboard side than for the sides of holds or the port side. There are no other significant differences.

Line diagrams are not appropriate, as the sample sizes are different.

**Tukey’s Procedure**

The Tukey’s MC procedure is constructed similarly to LSD, but with one critical difference. The distribution of the critical value is not t, but the studentized range distribution.

What is the studentized range distribution?

The Tukey Test and Interval

1)

2)

3)Keep in mind this is a pairwise difference test! Thus, all pairwise comparisons have an overall type I error rate of α%. In other words, the probability of observing one or more errors of all comparisons is α%.

Example:

**Student-Newman-Keuls Procedure**

The SNK MC procedure is constructed similarly to Tukey’s, but with a slight modification. We must consider the number of “steps” apart the means are when computing the margin of error. There will be a unique SNK difference for each number of steps apart.

The SNK Test and Interval

1)

2)

3)Keep in mind this is a pairwise difference test! Thus, all pairwise comparisons have an overall type I error rate of α%. In other words, the probability of observing one or more errors of all comparisons is α%.

Example:

**Dunnett’s Procedure: Comparison with a Control**

Dunnett’s procedure is different from all the other procedures we’ve considered. This procedure computes simultaneous comparisons with a control. Often there is one level of a treatment that can be considered a control. For example, we want to compare each dose of a drug back to the placebo or we want to compare each exercise method back to the group assigned no exercise program. In any case involving a control, we will use Dunnett’s procedure.

The Dunnett Test and Interval

1)

2)

3)

4)

5)The experimentwise error rate is controlled for comparison of each treatment level back to the control. This does not control for all-pairwise comparisons of the treatments!

Example:

**Scheffe’s S Method**

This is a more general simultaneous method. It should be thought of as distinct from the all-pairwise or controls MC methods. The Scheffe method allows for any number of arbitrary comparisons. Thus, you could write a set of contrasts to compare means to answer any conceivable set of research questions. The Scheffe procedure WILL control the type I error for any set of contrasts you could come up with. Thus, when a particular kind of set is not of interest, it is often wise to default to Scheffe S method to control your error.

1)

2)

3)

4)

5)

Example: See handout on simultaneous contrasts and make hand calculations here.

In summary, if we want to compare all the means, we use:

If we have a control, we use:

If we are setting up simultaneous contrasts of any number and type, we use: