Lab Tutorial 5: Controlling Type I Error - Multiple Comparisons

Chris M. Fiacconi

Multiple Comparisons: Illustrating the Problem

In this tutorial, we will discuss the multiple comparisons problem, namely, that when conducting multiple tests on a dataset, the probability of committing at least one Type I error can be substantially greater than the nominal value of .05. In other words, when conducting multiple tests each with a $\alpha_{PC} = .05$, the experiment or *family-wise* Type I Error rate, α_{FW}, is likely to exceed .05. Specifically, the probability of committing at least one Type I Error when testing C comparisons each with $\alpha_{PC} = .05$ can be computed as:

$$\alpha_{FW} = 1 - (1 - \alpha_{PC})^C$$

So if we conduct $C = 3$ tests, our $\alpha_{FW} = 1 - (1 - .05)^3 = .143$, which is substantially higher than the $\alpha_{PC} = .05$. As C increases, so too does our α_{FW}. One way of controlling this Type I Error inflation is to choose an α_{PC} such that, given C tests, will yield a $\alpha_{FW} = .05$. This adjusted α_{PC} can be calculated as:

$$\alpha_{PC} = 1 - \sqrt[C]{1 - \alpha_{FW}}$$

Based on the above equation, if we wished to test $C = 3$ contrasts, we would need to set $\alpha_{PC} = .017$ to maintain a $\alpha_{FW} = .05$.

Planned Vs. Post-Hoc Comparisons

Another questionable practice that can result in an increase in Type I Errors is inspecting the data *post-hoc* and then choosing which comparisons to perform. This is problematic because often the comparisons that one would choose to perform in this situation involve the largest group differences (to maximize the likelihood of obtaining a significant effect). Assuming the null hypothesis is true, always comparing the groups with the largest difference will result in a $\alpha_{PC} > .05$. Therefore, we need different approaches to controlling α_{FW} when the comparisons are planned (decided before looking at the data) vs. post-hoc (decided after looking at the data).

Controlling Type I Error - Planned Comparisons

Bonferroni Adjustment

When multiple comparisons are planned in advance of examining the pattern of means in the data, the most common method of controlling α_{FW} is Bonferroni's adjustment procedure. This approach requires that you simply divide the α_{PC} by the number of comparisons, C, to get the new α_{PC} that ensures $\alpha_{FW} = .05$. So if we are planning on performing $C = 3$ comparisons, our new $\alpha_{PC} = .05/3 = .0167$. The p-value associated with each comparison must now be $< .0167$ in order to qualify as statistically significant.

Note however, that we have the flexibility to divide up the original α_{PC} into three non-equal pieces. If the first comparison were of particular interest, and we wanted to maximize power, we could chose an $\alpha = .03$ for this comparison and use $\alpha = .01$ for the other two comparisons so long as they sum to .05. If we choose this approach, however, we have to decide how to divide up the original .05 **before** we look at the data.

Let's go through an example using the data from Example #1 in the previous tutorial:

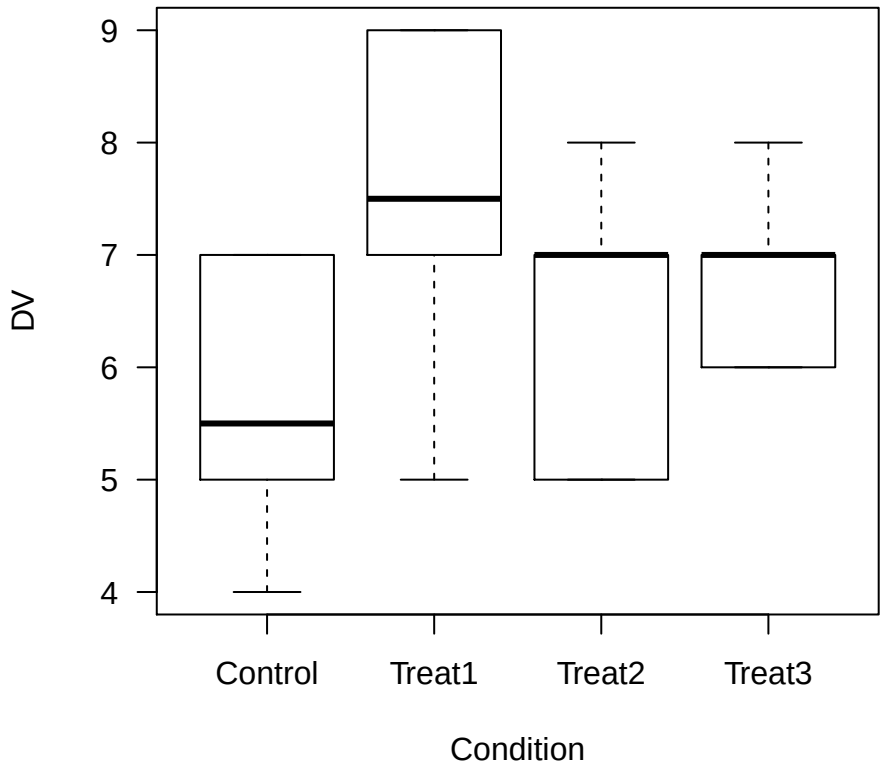

Figure 1: Boxplot of data derived from MDK Chapter 4 exercise 11

```
ex.data<-read.table(file="Ch4E11.txt",header=T)
boxplot(dv~cond,las=1,xlab="Condition",ylab="DV",data=ex.data)
```

In the previous tutorial, we went through an example in which we tested whether the average of the Treatment conditions was greater than the Control condition using a linear contrast. Let's say that prior to collecting these data, we also wanted to know whether each Treatment condition was greater than the Control condition. We could in principle conduct three t-tests to answer this question, but let's instead perform three linear contrasts of the following form:

$$\psi_1 = \mu_2 - \mu_1$$

$$\psi_2 = \mu_3 - \mu_1$$

$$\psi_3 = \mu_4 - \mu_1$$

By using linear contrasts rather than t-tests, our estimate of the population error variance is more precise, because we're pooling the error variance from all four conditions in each contrast, rather than using only the error variance from the two conditions we're comparing. Let's use the **linear.comparison** command to perform all three contrasts simultaneously:

```
source("lin_comp.R") # load in linear.comparison function

alpha.pc<-.05
n.comp<-3

adj.alpha<-alpha.pc/n.comp # Calculate new per-comparison alpha using Bonferroni adjustment
print(round(adj.alpha,digits=3))
```

```
## [1] 0.017
```

```
# Define contrast weights
c1<-c(-1,1,0,0)
c2<-c(-1,0,1,0)
c3<-c(-1,0,0,1)

my.contrasts<-list(c1,c2,c3) # Combine all contrasts into a list

linear.comparison(y=ex.data$dv,group=ex.data$cond,c.weights=my.contrasts,var.equal=TRUE)
```

```
## [1] "computing linear comparisons assuming equal variances among groups"
## [1] "C 1: F=6.914, t=2.630, p=0.016, psi=1.833, CI=(0.181,3.486), adj.CI= (0.012,3.655)"
## [1] "C 2: F=1.429, t=1.195, p=0.246, psi=0.833, CI=(-0.633,2.300), adj.CI= (-0.988,2.655)"
## [1] "C 3: F=2.800, t=1.673, p=0.110, psi=1.167, CI=(-0.048,2.381), adj.CI= (-0.655,2.988)"
```

Our new $\alpha_{PC} = .017$, so only p-values less than .017 are now considered statistically significant. You can see from the output that only the first comparison, $\psi_1 = \mu_2 - \mu_1$, $F(1, 20) = 6.91, p = .016, \psi_1 = 1.833$, was significant ($p < .017$) according to our new adjusted criterion for significance.

The Bonferroni procedure generally does a good job of ensuring that $\alpha_{FW} = .05$ for multiple planned comparisons when the group variances are equal. With unequal variances, the procedure is slightly different. See your textbook for details.

It should also be mentioned that some view adjusting α_{PC} for planned comparisons using the Bonferroni approach as too conservative. Remember, the smaller α_{PC} becomes, the less power we have to detect true population differences. For this reason, some researchers have argued that when a small number of comparisons (e.g., ≤ 3) are **planned in advance**, leaving $\alpha_{PC} = .05$ is justified. Note that this more liberal approach does necessarily mean that $\alpha_{FW} > .05$, but some argue that the corresponding increase in power more than offsets this increase in family-wise Type I Error. You'll see once we discuss factorial ANOVAs that often times researchers do indeed conduct 3 or more planned comparisons with $\alpha_{PC} = .05$.

Simultaneous Confidence Intervals

We have seen previously that there is a close relationship between 95% CIs and p-values, such that if a 95% CI on the difference between two population means does **not** include zero, then p is necessarily <.05 for that comparison. By definition, such a 95% CI will capture the true population difference between these means 19 times out of 20. However, this definition of a 95% CI is directly analogous to the situation where $\alpha_{PC} = .05$. When we are performing multiple comparisons, C, it is possible to instead compute a set of 95% CIs such that 95% of the time all C confidence intervals will contain the difference in their respective population means. These CIs are known as **simultaneous confidence intervals** and are directly analogous to situations where α_{PC} is modified to maintain $\alpha_{FW} = .05$. When using the Bonferroni adjustment to control α_{FW}, the formula for a 95% simultaneous confidence interval for multiple linear contrasts is:

$$\psi \pm \sqrt{F_{.05/C;1;N-k}} \sqrt{MS_W \sum_{j=1}^{k} (c_j^2/n_j)}$$

Fortunately, you don't need to calculate these intervals by hand, as the **linear.comparison** function outputs them for you. The 95% simultaneous CIs are given by the *adj.CI* value in the output.

95% simultaneous CIs can also be calculated from other multiple comparison procedures as well, including those designed to hand post-hoc comparisons. As we discuss the other methods of controlling α_{FW}, we will illustrate how such confidence intervals can be computed.

Post-Hoc Comparisons

So far, we have focused on how to control α_{FW} in situations where multiple planned comparisons are to be conducted. Although it is always advisable to develop *a priori* comparisons of interest, it is sometimes informative to test multiple comparisons after peeking at the data. This approach is only permissible if one adopts the appropriate procedures for controlling the α_{FW}. We will discuss how to conduct these post-hoc comparisons both in situations where all pairwise comparisons are of interest, and when multiple complex comparisons are desired.

All Pairwise Comparisons - Tukey HSD

In many situations, a researcher may want to examine all pairwise comparisons to locate the differences among population means. Often, the interest in all pairwise comparisons is piqued when the initial omnibus ANOVA test is significant. Recall that a significant omnibus ANOVA only indicates that the population means under consideration are not all equal. It does not specify which means are different. Therefore, pairwise comparisons are subsequently conducted to hone in on the specific group differences.

As discussed in Lab Tutorial 3, a common test used to evaluate all possible pairwise comparisons is Tukey's HSD procedure. This procedure allows the researcher to examine all pairwise contrasts in the data while maintaining $\alpha_{FW} = .05$. Tukey's HSD is based on the sampling distribution of the maximum pairwise difference among a set of pairwise contrasts. So under the null hypothesis that all population means are equal, one might ask, "How big should we expect the maximum difference among C pairwise comparisons to be based on random sampling error alone?" The F-value for the maximum pairwise difference is calculated as:

$$F_{pmax} = \frac{n(\overline{Y}_{max} - \overline{Y}_{min})}{2MS_W}$$

The distribution of this F-statistic is known as the *studentized range* distribution and is represented using the letter q where $q = \sqrt{2F_{pmax}}$. So for each pairwise comparison, you would convert the obtained F-value to a q-value using the above equation, and then compare this number to a critical q-value. This critical q-value is chosen to maintain $\alpha_{FW} = .05$. Note that as shown in Table 5.6 in your textbook, the critical value of q increases rapidly as the number of pairwise comparisons increases. Thus, the more pairwise comparisons you perform, the harder it becomes to reject the null hypothesis that no population group difference exists.

In *R* the Tukey HSD test is very straightforward to conduct. Simply perform a one-way ANOVA and save the output to an *aov* object. This *aov* object serves as the input argument to the **TukeyHSD** command:

```
ex.aov<-aov(dv~cond,data=ex.data) # Perform ANOVA to create aov object
TukeyHSD(ex.aov) # Pass aov object to TukeyHSD command
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = dv ~ cond, data = ex.data)
##
## $cond
##                      diff        lwr      upr     p adj
## Treat1-Control  1.8333333 -0.1181317 3.784798 0.0703035
## Treat2-Control  0.8333333 -1.1181317 2.784798 0.6367716
## Treat3-Control  1.1666667 -0.7847983 3.118132 0.3630075
## Treat2-Treat1  -1.0000000 -2.9514650 0.951465 0.4937164
## Treat3-Treat1  -0.6666667 -2.6181317 1.284798 0.7752460
## Treat3-Treat2   0.3333333 -1.6181317 2.284798 0.9630632
```

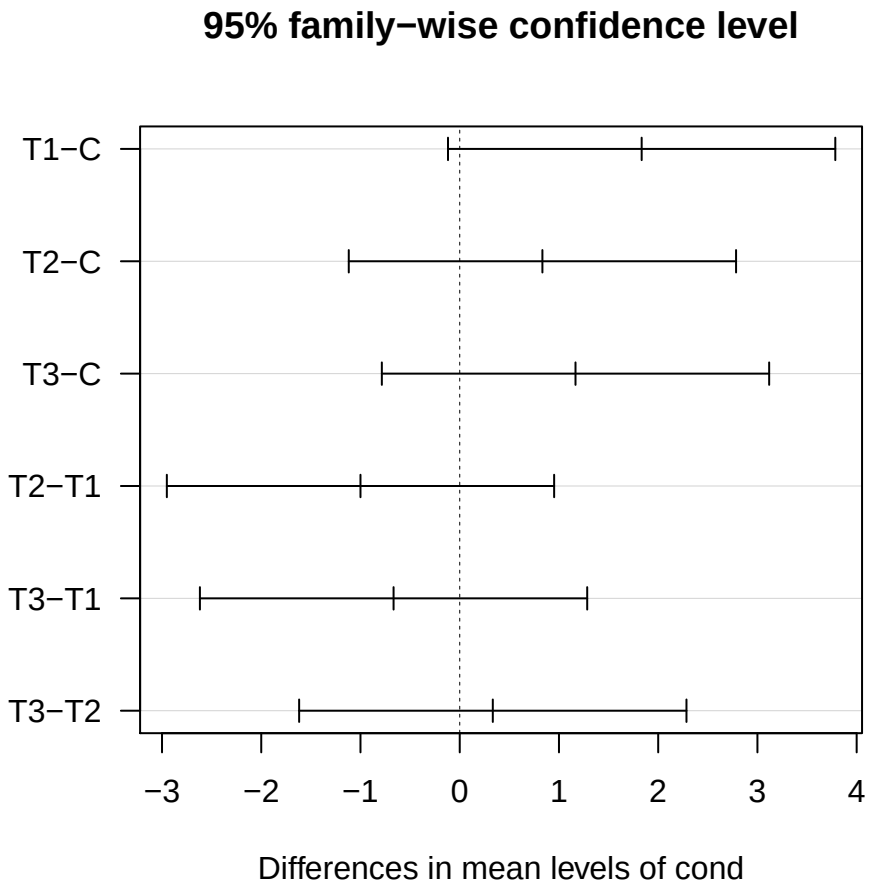

Figure 2: 95% Simultaneous Confidence Intervals based on Tukey's HSD pairwise comparison procedure

Note that it is not necessary to report the results of the ANOVA beforehand. If all pairwise comparisons are of interest, it is perfectly reasonable to jump right ahead and use the **TukeyHSD** function.

Simultaneous 95% CIs that correspond to pairwise comparisons using the Tukey HSD procedure can be calculated as follows:

$$(\bar{Y}_g - \bar{Y}_h) \pm (q_{.05;k,N-k}/\sqrt{2})\sqrt{MS_W(\frac{1}{n_g} + \frac{1}{n_h})}$$

Fortunately, these simultaneous 95% CIs are included in the output. The lower and upper limits of these CIs are indicated by *lwr* and *upr*, respectively. You can also plot these simulataneous 95% CIs with the following code:

```
plot(TukeyHSD(ex.aov),yaxt="n")
axis(side=2,at=c(6,5,4,3,2,1),labels=c("T1-C","T2-C","T3-C","T2-T1","T3-T1","T3-T2"),las=1)
```

It appears as though the first pairwise comparison (Treat1-Control) is not significant using Tukey's HSD test, but it was significant when we performed a linear contrast comparing these two groups above. Why might this be?

Complex Comparisons - Scheffe's Method

Tukey's HSD maintains $\alpha_{FW} = .05$, but is appropriate only when all post-hoc comparisons of interest are pairwise. When some of the comparisons involve complex linear contrasts, Scheffe's Method is used. Similar

to Tukey's HSD, Scheffe's Method requires us to compute a new critical value of our test statistic (in this case F) to which we compare our observed F-statistic. This new critical value of F is known as $F_{Scheffe}$:

$$F_{Scheffe} = (k-1) \times F_{\alpha FW}(df1 = k-1, df2 = N-k)$$

We can calculate $F_{Scheffe}$ in R for a study with $k = 4$ groups ($n = 6$ each) using the following commands:

```
alpha.fw<-.05
df.num<-4-1
df.denom<-24-4

F<-qf(1-alpha.fw,df1=df.num,df2=df.denom)
F.scheffe<-(4-1)*F
print(round(F.scheffe,digits=3))
```

```
## [1] 9.295
```

So for any linear contrast, ψ, if the observed value of $F > 9.295$, we can reject the null hypothesis. Note that you can now test as many post-hoc linear contrasts as you like so long as you use this critical value of F for each contrast.

Simultaneous 95% CIs that correspond to complex linear contrasts using the Scheffe Method can be calculated as follows:

$$\psi \pm \sqrt{(k-1)F_{.05;k-1;N-k}} \sqrt{MS_W \sum_{j=1}^{k}(c_j^2/n_j)}$$

Let's suppose that after running the study and plotting the means, I decide to test whether the Treat1 condition yields better results than the other two treatment conditions, Treat2 and Treat3. Furthermore, I want to test whether the Treat1 condition is different than the mean of the other three conditions. Let's use R to do this with the **linear.comparison** function together with the Scheffe Method of evaluating each linear contrast:

```
print(F.scheffe)
```

```
## [1] 9.295174
```

```
c1<-c(0,2,-1,-1) # Define contrast weights for comparison 1
c2<-c(-1,3,-1,-1) # Define contrast weights for comparison 2

phoc.contrasts<-list(c1,c2) # Combine contrast weights into a single list
linear.comparison(y=ex.data$dv,group=ex.data$cond,c.weights=phoc.contrasts,var.equal=TRUE)
```

```
## [1] "computing linear comparisons assuming equal variances among groups"
## [1] "C 1: F=1.905, t=1.380, p=0.183, psi=1.667, CI=(-1.192,4.525), adj.CI= (-1.260,4.593)"
## [1] "C 2: F=4.200, t=2.049, p=0.054, psi=3.500, CI=(-0.692,7.692), adj.CI= (-0.638,7.638)"
```

The obtained F-statistic for both contrasts $< F_{Scheffe} = 9.295$, so we fail to reject the null hypothesis for each test.

Another important feature of Scheffe's Method is that it is *mutually consistent* with the omnibus F-test. That is, if the omnibus F-test comparing all group means is significant, there will always be at least one contrast that is also significant. In contrast, if the omnibus F-test is not significant, then there are no contrasts that will be significant. This property of Scheffe's Method differs from Tukey's HSD procedure because with the latter, it is possible to reject the null hypothesis from the omnibus F-test and *not* find any significant pairwise comparisons with Tukey's HSD test.