Notes on Maxwell & Delaney

PSY710

5 Chapter 5 - Multiple Comparisons of Means

5.1 Inflation of Type I Error Rate

When conducting a statistical test, we typically set $\alpha = .05$ or $\alpha = .01$ so that the probability of making a Type I error is .05 or .01. Suppose, however, we conduct 100 such tests. Further suppose that the null hypothesis for each test is, in fact, true. Although α for each individual test may be .05, the probability of making at least one Type I error *across the entire set* of 100 tests is much greater than .05. Why? Because the more tests we do, the greater the chances of making an error. More precisely, the probability of making *at least* one Type I error is

$$P(\text{at least one Type I error}) = \alpha_{FW} = 1 - (1 - \alpha_{PC})^C \tag{1}$$

where C is the number of tests performed[1]. α_{PC} is the *per comparison* Type I error rate; it represents the probability of making a Type I error for each test. α_{FW}, on the other hand, is the *familywise* Type I error rate, and it represents the probability of making a Type I error across the entire family, or set, of tests. For the one-way designs we are considering, α_{FW} also equals the α level for the entire experiment, or the *experimentwise* error rate (α_{EW}). For the current example, $\alpha = .05$, $C = 100$, and so $\alpha_{FW} = 0.994079$, which means that it is very likely that we would make at least one Type I error in our set of 100 tests. Here, in a nutshell, is the problem of conducting multiple tests of group means: the probability of making a Type I error increases with the number of tests. If the number of tests, C, is large, then it becomes very likely that we will make a Type I error.

When we are conducting multiple tests on group means, we generally want to minimize Type I errors across the entire experiment, and so we need some way of maintaining α_{EW} at some reasonably low level (e.g., .05). One obvious way of controlling α_{EW} is to rearrange Equation 1 to calculate the α_{PC} that is required for a given α_{FW} and C:

$$\alpha_{PC} = 1 - (1 - \alpha_{FW})^{1/C} \tag{2}$$

According to Equation 2, when $C = 100$ and we want $\alpha_{FW} = .05$, we must set α_{PC} to .0005128.

5.2 Planned vs. Post-hoc Comparisons

I will compare the group means with a linear contrast that assumes equal variance across all groups. Suppose I conduct an experiment that compares the scores of subjects randomly assigned to eight different groups. After inspecting the data, shown in Figure 1, I decide to compare the means of groups 4 and 7 because the difference between those groups looks fairly large. I can do the test two ways. First I can simply compare the groups using a t test assuming equal group variances. In the following commands, notice how I use the `subset` command to extract the data from the two groups, and then use R's function `t.test`. I will set $\alpha = .05$. The null hypothesis is that the groups are equal; the alternative is that they differ.

```
> levels(g)

[1] "g1" "g2" "g3" "g4" "g5" "g6" "g7" "g8"
```

[1] Technically, Equation 1 is correct only if the tests form an orthogonal set and the sample sizes for each group are large.

```
> y.4<-subset(y,g=="g4") # get scores for group 4
> y.7<-subset(y,g=="g7") # get scores for group 7
> t.test(y.4,y.7,var.equal=TRUE) # do t-test assuming equal variances

        Two Sample t-test

data:  y.4 and y.7
t = 4.165, df = 18, p-value = 0.0005813
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 13.84 42.01
sample estimates:
mean of x mean of y
   111.33     83.41
```

The results are significant ($t = 4.16$, $df = 18$, $p = 0.00058$), so I reject the null hypothesis of no difference between groups 4 and 7.

In the second method, I will compare the groups using a linear contrast, again assuming equal variances across groups. One advantage of this method is it uses *all* of the groups to derive an estimate of the population error variance, whereas `t.test` only uses data from the two groups being compared. Not only is the estimated error variance likely to be more accurate, but the test will have many more degrees of freedom in the denominator and therefore be more powerful. As before, $\alpha = .05$ and the null hypothesis is that the groups are equal. (Note the double brackets that I use to read the results stored in `c.4vs7`).

```
> lc.source<-url("http://psycserv.mcmaster.ca/bennett/psy710/Rscripts/linear_contrast_v2.R")
> source(lc.source)

[1] "loading function linear.comparison"

> close(getConnection(lc.source));
> my.contrast<-list(c(0,0,0,1,0,0,-1,0) );
> c.4vs7 <- linear.comparison(y,g,c.weights=my.contrast )

[1] "computing linear comparisons assuming equal variances among groups"
[1] "C 1: F=9.915, t=3.149, p=0.002, psi=27.924, CI=(14.560,41.287), adj.CI= (10.245,45.602)"

> c.4vs7[[1]]$F

[1] 9.915

> c.4vs7[[1]]$t

[1] 3.149

> c.4vs7[[1]]$p.2tailed

[1] 0.002387
```

Again, the comparison between the two groups is significant ($t = 3.1487$, $df = 72$, $p = .002387$).

Finally, for completeness, I will do the comparison using the `lm()` command:

```
> newG <- g; # copy grouping factor
> contrasts(newG) <-  c(0,0,0,1,0,0,-1,0) # link contrast weights with newG
> newG.lm.01 <- lm(y~newG)
> summary(newG.lm.01) # print coefficients & t-tests
```

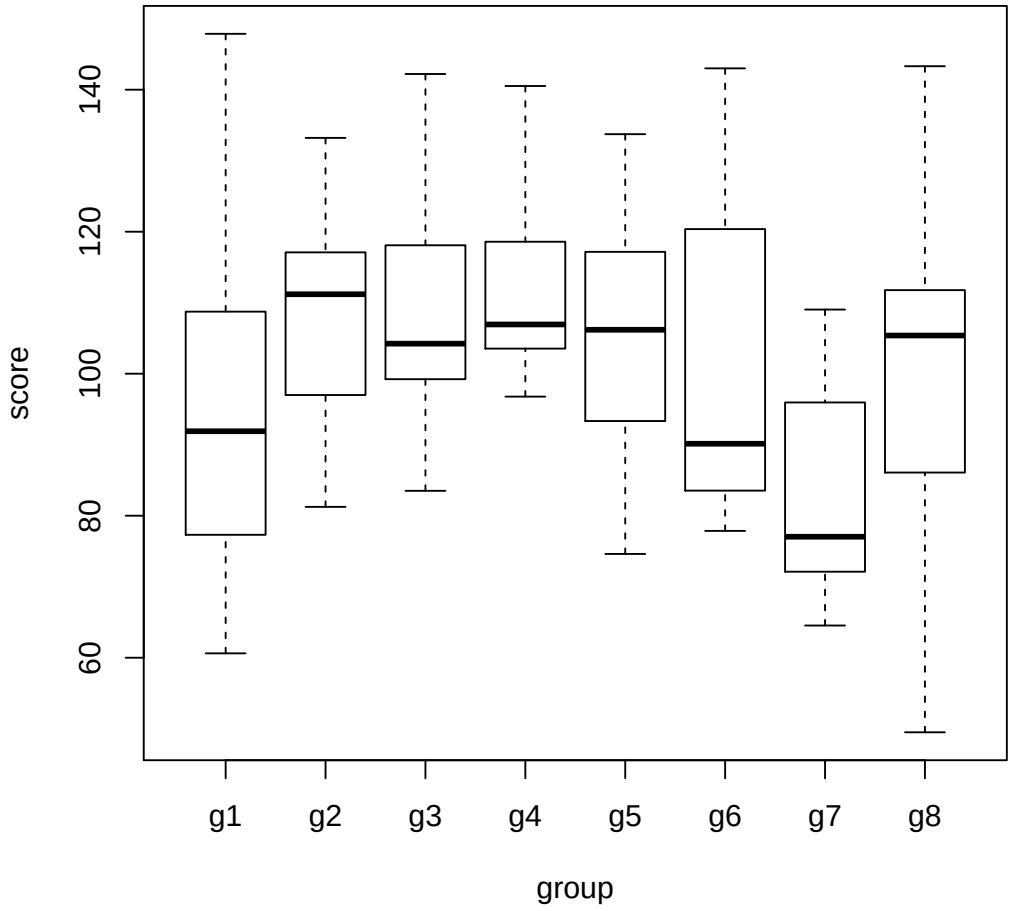

Figure 1: Eight sets of data

```
Call:
lm(formula = y ~ newG)

Residuals:
   Min     1Q Median     3Q    Max
-51.20 -12.17  -3.41  11.59  52.23

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  101.716      2.217   45.88   <2e-16 ***
newG1         13.962      4.434    3.15   0.0024 **
newG2          7.417      6.271    1.18   0.2408
newG3         -9.100      6.271   -1.45   0.1511
newG4          5.596      6.271    0.89   0.3751
newG5          0.149      6.271    0.02   0.9811
newG6          3.590      6.271    0.57   0.5688
newG7          0.573      6.271    0.09   0.9274
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 19.8 on 72 degrees of freedom
Multiple R-squared:  0.168,        Adjusted R-squared:  0.0873
F-statistic: 2.08 on 7 and 72 DF,  p-value: 0.0567
```

The first coefficient after the intercept, `newG1`, is the one we are interested in. The linear contrast is significant ($t(72) = 3.15$, $p = 0.0024$).

Question: What was wrong with the preceding analyses?

Well, notice that I did something a bit odd here: I first looked at the data and *then* decided what groups I would compare: I set $\alpha = .05$, and went off on my merry way. And that is where I made my mistake: the true α will be significantly greater than the nominal α. Why? Because when I looked at the data, it was as though I implicitly compared all of the groups and then decided to calculate the p-value for the largest value of t (or F): I compared the groups that *looked different*. Obviously, this pre-screening of the data will increase the probability of making a Type I error.

Comparing groups after looking at the data is what is known as a *post-hoc* comparison. *Planned* comparisons, on the other hand, are tests which have been planned before looking at the data. What is needed is a method for controlling Type I error rates for post-hoc comparisons.

5.3 Multiple Planned Comparisons

5.3.1 Bonferroni adjustment

Consider the situation where you are doing C planned comparisons and you want to have a familywise Type I error rate of $\alpha_{FW} = .05$. What should α_{PC} be?

A common method of controlling familywise Type I errors is to simply divide the per-comparison α by the number of comparisons:

$$\alpha_{PC} = \alpha_{FW}/C \tag{3}$$

This procedure for adjusting α_{PC} is known as the **Bonferroni adjustment**. It also is referred to as **Dunn's procedure**. If $\alpha_{FW} = .05$ and $C = 4$, then $\alpha_{PC} = .0125$. Therefore, in this case we would reject the null hypothesis for any tests whose p-value was $\leq .0125$. By using the Bonferroni adjustment,

$$\alpha_{FW} \leq C\alpha_{PC} \tag{4}$$

In this particular case ($C = 4$ and $\alpha_{PC} = .0125$), the true α_{FW} will be less than or equal to .05.

The Bonferroni inequality shown in Equation 4 is true only for a set of orthogonal contrasts. If the contrasts are not orthogonal, then the true value of α_{FW} is lower than its nominal value. In other words, the Bonferroni adjustment maintains $\alpha_{FW} \leq .05$ for orthogonal and non-orthogonal contrasts, but it is more conservative for non-orthogonal contrasts.

5.3.2 Holm's sequential Bonferroni test

Holm (1979) described a simple modification of the Bonferroni test that can increase power significantly. Assume that we are doing C comparisons that result in C statistics (e.g., C values of t or F) and that we want a familywise Type I error rate of α_{FW}. The first step is to rank order the *absolute values* of the statistics from highest to lowest. (N.B. If the sample sizes are not equal, then the various comparisons are ranked in terms of their p values, from lowest to highest.) Then we evaluate the first comparison at the α_{FW}/C level of significance, the second comparison at the $\alpha_{FW}/(C-1)$ level of significance, the third comparison at the $\alpha_{FW}/(C-2)$, and so on. The procedure stops when a comparison is not significant. Holm (1979) showed that this procedure, like the Bonferroni method, ensures that the true familywise Type I error rate is less than the nominal value of α_{FW}. However, Holm's method is more powerful than the Bonferroni procedure because the critical value of α increases starting with the second comparison. Kirk (1995, pages 142-144) discusses Holm's method and provides examples of how to use it.

5.3.3 setting α_{FW}

What value of α_{FW} should we use? Generally, α_{FW} is set to one of the two standard criteria (i.e., .05 or .01). However, in the case where you want to do a small number of multiple planned *orthogonal* contrasts, some have argued that a larger α_{FW} is justified (Keppel, 1982; Kirk, 1995). The basis of this argument is that the Bonferroni adjustment has the effect of reducing the power of each comparison, and that this reduction is too great a penalty in cases where an experimenter wants to do a relatively small number of planned comparisons. In fact, some have argued that the Bonferroni adjustment is unnecessary if you are doing $a-1$ planned comparisons, where a is the number of groups. There is some disagreement on the issue of whether the planned comparisons should be orthogonal.

My own feeling on this issue is that if you are doing a small number of planned, orthogonal comparisons, then it probably is OK to set $\alpha_{PC} = .05$. Notice that by doing so we will allow α_{FW} to increase beyond .05. Why do I think this inflation of α_{FW} is OK? Consider a case where we are analyzing a crossed, two-factor experiment. In other words, there are two experimental factors, and each level of factor A is crossed, or combined with, each level of factor B. As we will see later on this term, an analysis of such a design includes a test for the effects of factor A, of factor B, and what is called the AB interaction. Each test constitutes a *family* of tests, and $\alpha_{FW} = .05$. However, because there are three families of tests, and because each family is orthogonal to the others, $\alpha_{EW} = .15$. The point is that we often tolerate experiment-wise Type I error rates that are greater than .05 when there are two or more experimental factors. I see no reason, therefore, to get too upset when somebody analyzes a set of five means in a one-way design using three orthogonal contrasts with $\alpha_{PC} = .05$. However, it is important for *you* to understand that there is a trade off between Type I and Type II error rates. If you keep a tight lid on Type I experiment-wise error rates, then the power of each of your individual tests declines and you are more likely to make a Type II error. If you keep α_{PC} at .05, the power of each test improves but the experiment-wise Type I error increases, perhaps substantially.

5.3.4 simultaneous confidence intervals

Previously we have seen that there is a close connection between p-values and confidence intervals. So, if we are adjusting p-values to control α_{FW}, you might imagine that confidence intervals also should be adjusted whenever multiple, simultaneous intervals are constructed. You would be correct.

When we refer to a single 95% confidence interval, we mean that the interval contains the true population mean 95% of the time. When we refer to sets of simultaneous 95% confidence intervals, or adjusted 95%

confidence intervals, we mean that *all* of the intervals in the set will contain the population mean 95% of the time. Sets of two or more simultaneous confidence intervals will always be larger than individual confidence intervals.

The R routine that I have written for this class, `linear.comparison`, computes adjusted confidence intervals. As an example, we'll conduct three contrasts on the blood pressure data from Table 5.4 in your textbook.

```
> bp<-read.csv(url("http://psycserv.mcmaster.ca/bennett/psy710/datasets/maxwell_tab54.csv"))
> names(bp)

[1] "group"         "bloodPressure"

> blood.contrasts <- list(c(1,-1,0,0), c(1,0,-1,0), c(0,1,-1,0), c(1,1,1,-3)/3 );
> bp.results <- linear.comparison(bp$bloodPressure,bp$group,blood.contrasts,var.equal=TRUE)

[1] "computing linear comparisons assuming equal variances among groups"
[1] "C 1: F=1.562, t=1.250, p=0.226, psi=5.833, CI=(-6.106,17.773), adj.CI= (-6.976,18.643)"
[1] "C 2: F=0.510, t=0.714, p=0.483, psi=3.333, CI=(-5.185,11.852), adj.CI= (-9.476,16.143)"
[1] "C 3: F=0.287, t=-0.536, p=0.598, psi=-2.500, CI=(-13.521,8.521), adj.CI= (-15.310,10.310)"
[1] "C 4: F=9.823, t=3.134, p=0.005, psi=11.944, CI=(5.620,18.269), adj.CI= (1.485,22.403)"
```

Notice that I divided the weights of the last comparison by 3; this division will be important later.

Now let's compare the confidence intervals and *adjusted* confidence intervals for each comparison. These are all 95% confidence intervals because $\alpha = .05$. The confidence intervals in the variable `confinterval` are for the individual comparison. The confidence intervals in the variable `adj.confint` are adjusted, or simultaneous, confidence intervals. Note how the adjusted confidence intervals are always larger than the non-adjusted intervals. Also note that all of the adjusted confidence intervals include zero. What does this mean for the way we evaluate the null hypothesis for each comparison? If we evaluate the null hypothesis for each comparison using the unadjusted confidence interval, what is α_{FW}? What is α_{FW} if we use the adjusted confidence intervals?

The confidence intervals are for Ψ, which is defined as

$$\hat{\Psi} = \sum_{j=1}^{a} (c_j \bar{Y}_j)$$

Our fourth contrast, $c = (1/3, 1/3, 1/3, -1)$, yields

$$\Psi = (1/3)(\mu_1 + \mu_2 + \mu_3) - (1)\mu_4$$

which tests the null hypothesis that mean of group 4 is equal to the mean of the other group means. Now, a contrast like $c = (1, 1, 1, -3)$ also compares group 4 to the other 3 groups, but the value of Ψ is changed to

$$\Psi = (1)(\mu_1 + \mu_2 + \mu_3) - (3)\mu_4$$

which is three times the previous value. It turns out that the F and t tests are not affected by this manipulation, but the value of Ψ and the confidence intervals for Ψ are changed:

```
> blood.contrasts <- list(c(1,-1,0,0), c(1,0,-1,0), c(0,1,-1,0), c(1,1,1,-3) );
> bp.results <- linear.comparison(bp$bloodPressure,bp$group,blood.contrasts,var.equal=TRUE)

[1] "computing linear comparisons assuming equal variances among groups"
[1] "C 1: F=1.562, t=1.250, p=0.226, psi=5.833, CI=(-6.106,17.773), adj.CI= (-6.976,18.643)"
[1] "C 2: F=0.510, t=0.714, p=0.483, psi=3.333, CI=(-5.185,11.852), adj.CI= (-9.476,16.143)"
[1] "C 3: F=0.287, t=-0.536, p=0.598, psi=-2.500, CI=(-13.521,8.521), adj.CI= (-15.310,10.310)"
[1] "C 4: F=9.823, t=3.134, p=0.005, psi=35.833, CI=(16.860,54.806), adj.CI= (4.456,67.210)"

> bp.results[[4]]$contrast
```

```
[1]  1  1  1 -3

> bp.results[[4]]$alpha

[1] 0.05

> bp.results[[4]]$confinterval

[1] 16.86 54.81

> bp.results[[4]]$adj.confint

[1]  4.456 67.210
```

In this case, the confidence interval is for a value that corresponds to three times the difference between μ_4 and $(1/3)(\mu_1 + \mu_2 + \mu_3)$. So, make sure that the confidence interval is for the value that you really care about.

One more thing. The p-values returned by `linear.comparison` are not adjusted for multiple comparisons. However, it is rather simple to compare them to an adjusted criterion. If we want to set $\alpha_{FW} = .05$, we can use the Bonferroni adjustment to calculate the correct value of α for each comparison: $\alpha_{PC} = \alpha_{FW}/C = .05/3 = .0167$. By default, the p-values are listed in the output of `linear.comparison`. Manually listing the unadjusted p-values for each comparison is easy:

```
> bp.results[[1]]$p.2tailed

[1] 0.2258

> bp.results[[2]]$p.2tailed

[1] 0.4834

> bp.results[[3]]$p.2tailed

[1] 0.5981

> bp.results[[4]]$p.2tailed

[1] 0.005224
```

Only the fourth comparison is significant.

5.4 All Pairwise Contrasts

Sometimes you will be interested only in pairwise comparisons between group means. If you want to test just a few pairs of means, it often is sufficient to do multiple t-tests (or to use linear contrasts to compare pairs of groups) and to use the Bonferroni adjustment (or Holm's sequential method) to control α_{FW}. However, when you want to do many pairwise tests, or you decide to do multiple pairwise tests after inspecting the data (i.e., post-hoc pairwise tests), then you should use `TukeyHSD`. (N.B. Tukey HSD is the same as Tukey WSD). Tukey's HSD (for Honestly Significant Difference) is the optimal procedure for doing *all* pairwise comparisons. In R, the Tukey test is invoked by calling `TukeyHSD(my.aov)`, where `my.aov` is an object created by a call to `aov`. Here is an example, again using the data in Table 5.4:

```
> bp.aov<-aov(bloodPressure~group,data=bp)
> TukeyHSD(bp.aov)
```

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = bloodPressure ~ group, data = bp)

$group
       diff    lwr    upr  p adj
b-a  -5.833 -18.90  7.231 0.6038
c-a  -3.333 -16.40  9.731 0.8903
d-a -15.000 -28.06 -1.936 0.0209
c-b   2.500 -10.56 15.564 0.9493
d-b  -9.167 -22.23  3.898 0.2345
d-c -11.667 -24.73  1.398 0.0906
```

The output is easier to read by calling `TukeyHSD` slightly differently:

```
> TukeyHSD(bp.aov,ordered=TRUE)

  Tukey multiple comparisons of means
    95% family-wise confidence level
    factor levels have been ordered

Fit: aov(formula = bloodPressure ~ group, data = bp)

$group
      diff     lwr   upr  p adj
b-d  9.167  -3.898 22.23 0.2345
c-d 11.667  -1.398 24.73 0.0906
a-d 15.000   1.936 28.06 0.0209
c-b  2.500 -10.564 15.56 0.9493
a-b  5.833  -7.231 18.90 0.6038
a-c  3.333  -9.731 16.40 0.8903
```

The output is a list of pairwise comparisons. For each comparison, the output shows the difference, the confidence interval of the difference (95% by default), and a p-value that can be used to evaluate the null hypothesis that the group difference is zero. The confidence intervals and p-values are adjusted to take into account the multiple comparisons. Using `TukeyHSD` ensures that the familywise Type I error rate is controlled (.05, by default). The Tukey test assumes equal sample size in every group, and homogeneity of variance.

It is important to note that it is *not* necessary to obtain a significant omnibus F test before using the Tukey HSD procedure. In fact, requiring a significant omnibus test means that the actual α for the Tukey HSD test will be significantly lower than the nominal value. If you want to make all pairwise comparisons, then it is perfectly reasonable to skip the regular ANOVA and use the Tukey HSD procedure (Wilcox, 1987).

5.4.1 Modifications of Tukey HSD

The Tukey HSD procedure assumes equal sample sizes and constant variance across groups. If those assumptions are not valid, then the p values and confidence intervals calculated with the Tukey HSD procedure will be incorrect. When the constant variance assumption is valid but sample sizes are unequal, the Tukey-Kramer test is recommended. When the variances are heterogeneous and sample sizes are unequal, Dunnett's T3 test is recommended. The following code shows how to use the DTK package in R to perform these tests on the blood pressure data used in the previous section:

```
> # install.packages("DTK") # download package and install on computer
> library("DTK") # load package into workspace
> TK.test(x=bp$bloodPressure,f=bp$group,a=0.05) # Tukey-Kramer test

  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = x ~ f)

$f
       diff    lwr    upr  p adj
b-a  -5.833 -18.90  7.231 0.6038
c-a  -3.333 -16.40  9.731 0.8903
d-a -15.000 -28.06 -1.936 0.0209
c-b   2.500 -10.56 15.564 0.9493
d-b  -9.167 -22.23  3.898 0.2345
d-c -11.667 -24.73  1.398 0.0906

> DTK.test(x=bp$bloodPressure,f=bp$group,a=0.05) # Dunnett's T3 test

[[1]]
[1] 0.05

[[2]]
       Diff Lower CI Upper CI
b-a  -5.833   -26.95  15.2867
c-a  -3.333   -18.40  11.7357
d-a -15.000   -29.60  -0.4001
c-b   2.500   -17.00  21.9961
d-b  -9.167   -28.30   9.9692
d-c -11.667   -23.80   0.4659

> # following commands store result and then plot simultaneous confidence intervals:
> # tmp <- DTK.test(x=bp$bloodPressure,f=bp$group,a=0.05)
> # DTK.plot(tmp)
```

These tests, plus several others, are described in detail by Kirk (1995, pages 146-150).

5.5 Post-Hoc Comparisons

The final situation we will consider is the case where you want to perform post-hoc linear contrasts, some of which are not pairwise comparisons. In this situation, Tukey's HSD procedure is not appropriate because not all of the comparisons are pairwise. Neither is the Bonferroni procedure, because the contrasts are post-hoc, not planned. Instead, we should use Scheffé's method.

Scheffé's method allows us to perform multiple, complex linear contrasts after looking at the data, while maintaining control of the Type I error rate. The method of computing the linear contrasts are exactly the same as the one used for planned linear contrasts. The only difference is that the observed $F_{contrast}$ is compared to $F_{Scheffe}$

$$F_{Scheffe} = (a-1) \times F_{\alpha_{FW}}(df_1 = a-1;\ df_2 = N-a) \tag{5}$$

where a is the number of groups and N is the total number of subjects. Suppose we have 40 subjects ($N = 40$) divided among 4 groups ($a = 4$) and we want α_{FW} to equal .05. The value of $F_{Scheffe}$ is

```
> alpha.fw <- .05
> (4-1)*qf(1-alpha.fw,df1=4-1,df2=40-4)

[1] 8.599
```

The critical value of F is 8.599. Now we conduct any linear contrast on the group means - as many as we want. If $F_{contrast} \geq 8.599$, then the contrast is significant[2]. The familywise error rate is .05.

If the omnibus F test is significant, then there will be at least one contrast that is significant using Scheffé's method. However, if the omnibus test is not significant, then it is impossible to find a significant contrast using Scheffé's method. **Hence, Scheffé's method and the omnibus F test are mutually consistent.** Such is not the case with some other methods. For example, it is possible to reject the omnibus null hypothesis and still fail to find a significant pairwise difference using Tukey's HSD method. The opposite can also occur: Tukey's HSD method may find significant differences even when the omnibus F is not significant.

References

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6:65–70.

Keppel, G. (1982). *Design and analysis: A researcher's handbook*. Prentice Hall, Englewood Cliffs, NJ, 2nd edition.

Kirk, R. E. (1995). *Experimental design: Procedures for the behavioral sciences*. Brooks/Cole, 3rd edition.

Wilcox, R. R. (1987). New designs in analysis of variance. *Annual Review of Psychology*, 38:29–60.

[2]If you evaluate a comparison using the t statistic, simply convert it to F using the formula $F = t^2$, and then compare it to $F_{Scheffe}$.