Chapter 13

Multiple Comparisons

One-way ANOVA, as well as more complex variants, provides a test of an overall null hypothesis of the form $H_0 : \alpha_i = 0$ for all i vs. $H_1 :$ some $\alpha_i \neq 0$. If we obtain a small P value for this test, it provides evidence against H_0 and in favor of H_1. However, this overall test provides little information on whether particular groups are different. We now turn to statistical methods designed to compare pairs of groups for one-way ANOVA designs. These procedures allow comparisons to be made among all possible pairs of groups, or sometimes one group vs. all others, and are collectively called **multiple comparisons**. Although multiple comparisons are often conducted in association with ANOVA, they are in fact stand-alone procedures (Hsu 1996). There is no need to conduct an ANOVA before using these procedures, although SAS will generate an overall F test regardless. Moreover, significant differences between groups in multiple comparisons may not coincide with a significant overall F test, or vice versa.

13.1 Models for multiple comparisons

The statistical model for multiple comparisons is basically the one-way ANOVA model expressed in a different form. The one-way ANOVA model is

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}, \tag{13.1}$$

where μ is the grand mean, α_i is the deviation from the grand mean caused by the ith group, and $\epsilon_{ij} \sim N(0, \sigma^2)$. For multiple comparison procedures it

is common to define $\mu_i = \mu + \alpha_i$, and so the one-way model becomes

$$Y_{ij} = \mu_i + \epsilon_{ij}. \tag{13.2}$$

We can think of μ_i as the mean of the ith group, where there are a total groups.

Now consider two groups i and j in a study which have means μ_i and μ_j, where $i \neq j$. We will be interested in estimating the difference in the means of these two groups, $\mu_i - \mu_j$, and finding a confidence interval to accompany this estimate for all possible pairs of groups. We will also be interested in testing whether the means of the two groups are equal, namely $H_0 : \mu_i = \mu_j$ or equivalently $H_0 : \mu_i - \mu_j = 0$, again for all possible pairs of groups. For a study with a groups, this amounts to $a(a-1)/2$ pairs of groups. For example, if there are $a = 3$ groups there are $3(3-1)/2 = 3$ possible pairwise comparisons (groups 1-2, 2-3, and 1-3). There are multiple comparison methods that provide estimates, confidence intervals, and tests, while others provide only tests but have more statistical power. The basic purpose of these procedures is to statistically test which pairs of treatments are different, and provide some idea of the magnitude of the difference. We will examine three procedures in this category, known as **all possible pairwise comparisons**. The procedures are called Fisher's least significant difference, the Tukey procedure, and the Ryan-Einot-Gabriel-Welsch (REGW) procedure (Hsu 1996).

For experiments that have a clearly identifiable control group, it may be appropriate to compare each group with only the control. For example, suppose the control is a standard drug treatment for a disease. We may only be interested in treatments that give a significantly better (or maybe worse) result compared to the control, and are not interested in other comparisons among the treatments. For a study with a groups including the control, this amounts to $a - 1$ pairs of groups with the control. For example, if there are $a = 3$ groups with the first group ($i = 1$) the control, there are $3 - 1 = 2$ possible comparisons (groups 1-2 and 1-3). We will examine Dunnett's procedure in this category, known as **multiple comparisons with a control** (Hsu 1996).

13.2 Error rates in multiple comparisons

There are two error rates commonly used to describe multiple comparison procedures. One is the **per comparison** error rate, which is the Type I

error rate for a single test comparing a single pair of groups. This rate is like that used in other statistical tests we have encountered, where only a single test is considered. The second is the **experimentwise error rate**, or **EER**. **The EER is defined as the probability of one or more Type I errors (rejecting H_0 when it is true) in a set of comparisons.**

Why do we need two error rates? Multiple comparison procedures such as the ones mentioned above can involve a substantial number of statistical tests, one test for each pair of groups. For example, with $a = 5$ groups there would be $5(5-1)/2 = 10$ possible pairwise comparisons, while for $a = 10$ groups we would have $10(10-1)/2 = 45$ comparisons! Given this many comparisons and tests, it is quite possible that some pairs would yield a significant test result even if the null hypothesis were true, i.e., we would reject $H_0 : \mu_i = \mu_j$ for one or more pairs of groups, even though there is no difference between the groups. For example, suppose that the per comparison error rate is set at the typical $\alpha = 0.05$ value, which amounts to a 1 in 20 chance of rejecting H_0 when it is true. Given $a = 10$ and 45 total tests, we would expect to see a few significant test results just by chance. This difficulty has been called the **multiplicity problem** (Westfall et al. 1999).

To see the magnitude of the multiplicity problem, we can plot the EER for the least significant difference procedure, which controls the per comparison error rate but not the EER. Fig. 13.1 shows a plot of the EER vs. the number of groups or treatments (a). The least significant difference procedure is a t test that compares the means for each pair of groups, with each test conducted at the same α level, in this case $\alpha = 0.05$. We see that the EER, and the number of pairwise comparisons, increases rapidly with the number of groups. Thus, it becomes more likely that any significant differences reported among groups are in fact Type I errors. In contrast, methods designed to control the EER, such as the Tukey procedure, would maintain an EER of 0.05 regardless of the number of groups. These tests manage the EER by essentially reducing the per comparison error rate for each test. **The penalty of controlling the EER is a loss of power to detect differences among groups where they do exist.**

Multiple comparison procedures have been the subject of considerable controversy in the ecological and statistical literature. Several tests you may encounter in the literature, such as least significant difference, Fisher's protected least significant difference, Duncan's multiple range test, and the Student-Newman-Keuls test, were very popular because they gave significant results more often than competing methods. Unfortunately, these particular

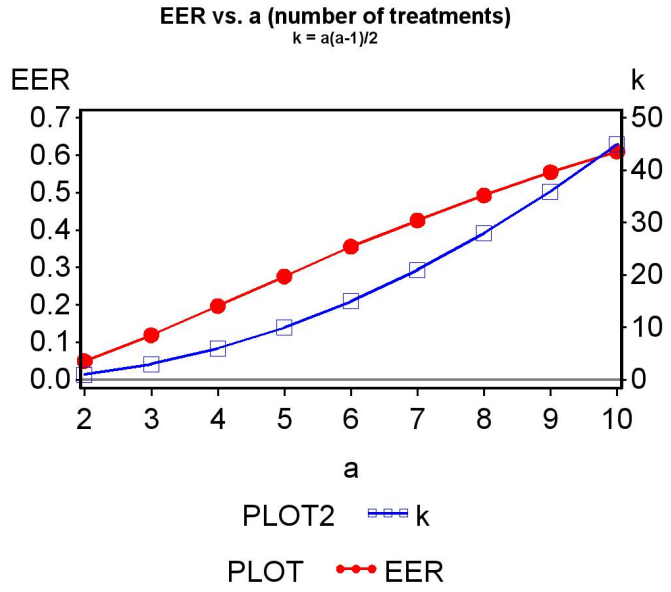

Figure 13.1: Plot of the experimentwise error rate vs. a, the number of treatments or groups, using $\alpha = 0.05$ for each comparison. Also shown is the number of pairwise comparisons ($k = a(a-1)/2$) vs. a.

tests do not control the experimentwise error rate (Day & Quinn 1989, Hsu 1996).

Another error rate that is becoming popular is the **false discovery rate** or **FDR** (Benjamini & Hochberg 1995).. **This is defined as the proportion of Type I errors in a set of comparisons.** Procedures that use the FDR have more power than those controlling the EER, but with more Type I errors. We will examine the rationale for FDR procedures later in the chapter.

13.3 All pairwise comparisons

This section examines three different methods for all pairwise comparisons among groups, the least significant difference, Tukey, and REGW methods. The least significant difference method does not control the EER, but is simple in form and a useful starting point. It provides estimates and confidence

intervals for $\mu_i - \mu_j$, the difference between the group means for any pair of groups, as well as a statistical test for $H_0 : \mu_i - \mu_j$. The Tukey procedure is similar to the least significant difference except that it controls the EER. We also examine the REGW method, an example of a **multiple range test**. Multiple range procedures only provide tests, not confidence intervals, but are more powerful procedures.

13.3.1 Least significant difference

We first develop confidence intervals and construct statistical tests for the least significant difference procedure, using methods similar to those in Chapter 9 and 10. For multiple comparisons, we are interested in estimating $\mu_i - \mu_j$ and finding a confidence interval for this quantity. It seems reasonable to use $\bar{Y}_i - \bar{Y}_j$ to estimate $\mu_i - \mu_j$, but what is the variance of this estimate? Using the rules for calculating the variance of a sum of random variables (Chapter 7), we have

$$Var[\bar{Y}_i - \bar{Y}_j] = Var[\bar{Y}_i] + (-1)^2 Var[\bar{Y}_j] = \sigma^2/n + \sigma^2/n = 2\sigma^2/n. \quad (13.3)$$

ANOVA provides an estimate of σ^2, namely MS_{within}, and so we can estimate the variance of $\bar{Y}_i - \bar{Y}_j$ using the quantity $2MS_{within}/n$, which has $a(n-1)$ degrees of freedom. Using these results, it can be shown that the quantity

$$\frac{(\bar{Y}_i - \bar{Y}_j) - (\mu_i - \mu_j)}{\sqrt{\frac{2MS_{within}}{n}}} \sim t_{a(n-1)}. \quad (13.4)$$

We use this quantity to first derive a confidence interval for $\mu_i - \mu_j$. Using Table T, we can find a value of $c_{\alpha,a(n-1)}$ for $a(n-1)$ degrees of freedom such that the following equation is true:

$$P\left[-c_{\alpha,a(n-1)} < \frac{(\bar{Y}_i - \bar{Y}_j) - (\mu_i - \mu_j)}{\sqrt{\frac{2MS_{within}}{n}}} < c_{\alpha,a(n-1)}\right] = 1 - \alpha. \quad (13.5)$$

Rearranging this equation, we obtain

$$P\left[\bar{Y}_i - \bar{Y}_j - c_{\alpha,a(n-1)}\sqrt{\frac{2MS_{within}}{n}} < \mu_i - \mu_j < \bar{Y}_i - \bar{Y}_j + c_{\alpha,a(n-1)}\sqrt{\frac{2MS_{within}}{n}}\right] = 1 - \alpha. \quad (13.6)$$

The confidence interval would therefore be the interval

$$\left(\bar{Y}_i - \bar{Y}_j - c_{\alpha,a(n-1)}\sqrt{\frac{2MS_{within}}{n}}, \bar{Y}_i - \bar{Y}_j + c_{\alpha,a(n-1)}\sqrt{\frac{2MS_{within}}{n}}\right). \quad (13.7)$$

The center of the confidence interval is located at $\bar{Y}_i - \bar{Y}_j$, the estimate of $\mu_i - \mu_j$. We will later illustrate how this interval is calculated in a SAS demo of the least significant difference procedure.

Now suppose we want to test $H_0 : \mu_i = \mu_j$ or equivalently $H_0 : \mu_i - \mu_j = 0$. Under H_0, the test statistic

$$T_s = \frac{(\bar{Y}_i - \bar{Y}_j) - 0}{\sqrt{\frac{2MS_{within}}{n}}} = \frac{(\bar{Y}_i - \bar{Y}_j)}{\sqrt{\frac{2MS_{within}}{n}}} \sim t_{a(n-1)}. \quad (13.8)$$

Using a Type I error rate of α, the acceptance region of the test would be the interval $(-c_{\alpha,a(n-1)}, c_{\alpha,a(n-1)})$, where $c_{\alpha,a(n-1)}$ is determined using Table T (see Chapter 10). We would reject H_0 if it falls on the edge or outside this interval.

We can rearrange the test given above into a different form, one that is commonly used for multiple comparisons. Recall that one would accept H_0 if T_s falls inside the acceptance region $(-c_{\alpha,a(n-1)}, c_{\alpha,a(n-1)})$, which implies

$$-c_{\alpha,a(n-1)} < \frac{(\bar{Y}_i - \bar{Y}_j)}{\sqrt{\frac{2MS_{within}}{n}}} < c_{\alpha,a(n-1)}. \quad (13.9)$$

We can rearrange this into the form

$$-c_{\alpha,a(n-1)}\sqrt{\frac{2MS_{within}}{n}} < \bar{Y}_i - \bar{Y}_j < c_{\alpha,a(n-1)}\sqrt{\frac{2MS_{within}}{n}}, \quad (13.10)$$

or

$$-LSD < \bar{Y}_i - \bar{Y}_j < LSD, \quad (13.11)$$

where

$$LSD = c_{\alpha,a(n-1)}\sqrt{\frac{2MS_{within}}{n}}. \quad (13.12)$$

The quantity LSD is called the least significant difference. We would accept H_0 if $\bar{Y}_i - \bar{Y}_j$ falls inside the interval $(-LSD, LSD)$, or equivalently if $|\bar{Y}_i - \bar{Y}_j| < LSD$. Conversely, we would reject H_0 if $|\bar{Y}_i - \bar{Y}_j| \geq LSD$. This

same rule applies to any pair of groups, because LSD would take the same value. Any pair of means that equals or exceeds this value is declared to be significantly different.

The confidence intervals we derived for $\mu_i - \mu_j$ can also be expressed in this format. In particular, the confidence interval would have the form

$$\left(\bar{Y}_i - \bar{Y}_j - LSD, \bar{Y}_i - \bar{Y}_j + LSD\right). \tag{13.13}$$

13.3.2 Least significant difference - SAS demo

Kneitel & Lessin (2010) studied the effect of eutrophication on vernal pools in California. They were interested in the effect of eutrophication (nutrient addition) on algae cover during the period the pools were filled with water, as well as vascular plant cover later in the season. Experimental pools were subjected to five different treatments: low, medium, high, and very high nutrient addition levels, and a control to which no nutrients were added. We will use a simplified data set from this study to illustrate the least significant difference procedure in SAS. We first examine the data involving algae cover. Algae cover was expressed as a percentage of the pool covered, and for data of this type it is common to transform the data. The data were first converted to a proportion by dividing the percentage by 100, then the arcsine-square root transformation applied (see Chapter 15). See the `data` step in the SAS program below.

The program is similar to our previous one-way ANOVA programs, with the addition of a `means` statement within `proc glm`:

```
means treat / t cldiff lines;
```

This statement requests a mean for each level of `treat`, the treatment variable (SAS Institute Inc. 2014). The `t` option requests the least significant difference procedure, because it is essentially a t test. The option `cldiff` requests 95% confidence intervals for $\mu_i - \mu_j$ for all pairs of groups, while `lines` generates a diagram that indicates which pairs of groups are significantly different at the $\alpha = 0.05$ level. See the full program listing and SAS output below.

According to the one-way ANOVA results, there was a highly significant difference among the nutrient treatments ($F_{4,20} = 4.76, P < 0.0073$). Confidence intervals for $\mu_i - \mu_j$ and $\mu_j - \mu_i$ are given for every pair of groups. For example, SAS gives a confidence interval for $\mu_{\text{medium}} - \mu_{\text{control}}$ as well as $\mu_{\text{control}} - \mu_{\text{medium}}$. Also shown in the output is the diagram generated by the

`lines` command. **Treatments with different letters are significantly different, while if they have the same letter they are not significantly different.** According to the letters, the very high, high, and medium treatments are significantly different from the low and control treatments, while there were no significant differences within these two groups. This lettering scheme can also be used to indicate significant differences among treatments within a graph (Fig. 13.2).

SAS Program

```
* Kneitel_2010_algae_lsd2.sas;
options pageno=1 linesize=80;
goptions reset=all;
title 'Multiple comparisons for algae cover';
title2 'Data from Kneitel and Lessin (2010)';
data kneitel;
        input treat $ richness total algae;
        * Apply transformations here;
        y = arsin(sqrt(algae/100));
        datalines;
Control   8   78    1
Control   5   84    7
Control  10  115   45
Control   7  200  100
Control   6   72   20
Low       8   73   15
Low       7  124   70
Low       8  116   50
Low       8   92    5
Low       7  138   60
Medium    7  124   85
Medium    8  116   80
Medium    8  145   60
Medium    6  154  100
Medium    7  129   90
High      6  134   95
High      7  138   95
High      8  103   70
High      8  119   75
High          6  132   80
VeryHigh  6  148   95
VeryHigh  5  134   95
VeryHigh  5  119  100
VeryHigh  5  117   90
VeryHigh  5  129   80
```

```
;
run;
* Print data set;
proc print data=kneitel;
run;
* Plot means, standard errors, and observations;
proc gplot data=kneitel;
        plot y*treat=1 / vaxis=axis1 haxis=axis1;
        symbol1 i=std1mjt v=star height=2 width=3;
        axis1 label=(height=2) value=(height=2) width=3 major=(width=2) minor=none;
run;
* One-way anova with comparisons;
proc glm data=kneitel;
        class treat;
        model y = treat;
        output out=resids p=pred r=resid;
        * LSD or Students t - only controls the per comparison error rate;
        means treat / t cldiff lines;
run;
goptions reset=all;
title "Diagnostic plots to check anova assumptions";
* Plot residuals vs. predicted values;
proc gplot data=resids;
        plot resid*pred=1 / vaxis=axis1 haxis=axis1;
        symbol1 v=star height=2 width=3;
        axis1 label=(height=2) value=(height=2) width=3 major=(width=2) minor=none;
run;
* Normal quantile plot of residuals;
proc univariate noprint data=resids;
        qqplot resid / normal waxis=3 height=4;
run;
quit;
```

SAS Output

Multiple comparisons for algae cover 1
Data from Kneitel and Lessin (2010)
15:51 Tuesday, July 3, 2012

Obs	treat	richness	total	algae	y
1	Control	8	78	1	0.10017
2	Control	5	84	7	0.26776
3	Control	10	115	45	0.73531
4	Control	7	200	100	1.57080
5	Control	6	72	20	0.46365
6	Low	8	73	15	0.39770
7	Low	7	124	70	0.99116
8	Low	8	116	50	0.78540
9	Low	8	92	5	0.22551
10	Low	7	138	60	0.88608
11	Medium	7	124	85	1.17310
12	Medium	8	116	80	1.10715
13	Medium	8	145	60	0.88608
14	Medium	6	154	100	1.57080
15	Medium	7	129	90	1.24905
16	High	6	134	95	1.34528
17	High	7	138	95	1.34528
18	High	8	103	70	0.99116
19	High	8	119	75	1.04720
20	High	6	132	80	1.10715
21	VeryHigh	6	148	95	1.34528
22	VeryHigh	5	134	95	1.34528
23	VeryHigh	5	119	100	1.57080
24	VeryHigh	5	117	90	1.24905
25	VeryHigh	5	129	80	1.10715

Multiple comparisons for algae cover 2
Data from Kneitel and Lessin (2010)
15:51 Tuesday, July 3, 2012

The GLM Procedure

Class Level Information

Class	Levels	Values

```
          treat             5    Control High Low Medium VeryHigh


                    Number of Observations Read          25
                    Number of Observations Used          25


                      Multiple comparisons for algae cover                    3
                      Data from Kneitel and Lessin (2010)
                                                15:51 Tuesday, July 3, 2012

                              The GLM Procedure

Dependent Variable: y

                                        Sum of
 Source                       DF        Squares    Mean Square   F Value   Pr > F

 Model                         4     2.13816313     0.53454078      4.76   0.0073

 Error                        20     2.24444069     0.11222203

 Corrected Total              24     4.38260382


              R-Square     Coeff Var      Root MSE        y Mean

              0.487875      33.68371      0.334996      0.994533


 Source                       DF      Type I SS    Mean Square   F Value   Pr > F

 treat                         4     2.13816313     0.53454078      4.76   0.0073


 Source                       DF    Type III SS    Mean Square   F Value   Pr > F

 treat                         4     2.13816313     0.53454078      4.76   0.0073


                      Multiple comparisons for algae cover                    4
                      Data from Kneitel and Lessin (2010)
                                                15:51 Tuesday, July 3, 2012

                              The GLM Procedure
```

t Tests (LSD) for y

NOTE: This test controls the Type I comparisonwise error rate, not the experimentwise error rate.

Alpha	0.05
Error Degrees of Freedom	20
Error Mean Square	0.112222
Critical Value of t	2.08596
Least Significant Difference	0.442

Comparisons significant at the 0.05 level are indicated by ***.

treat Comparison	Difference Between Means	95% Confidence Limits		
VeryHigh - Medium	0.1263	-0.3157	0.5682	
VeryHigh - High	0.1563	-0.2857	0.5983	
VeryHigh - Low	0.6663	0.2244	1.1083	***
VeryHigh - Control	0.6960	0.2540	1.1379	***
Medium - VeryHigh	-0.1263	-0.5682	0.3157	
Medium - High	0.0300	-0.4119	0.4720	
Medium - Low	0.5401	0.0981	0.9820	***
Medium - Control	0.5697	0.1277	1.0116	***
High - VeryHigh	-0.1563	-0.5983	0.2857	
High - Medium	-0.0300	-0.4720	0.4119	
High - Low	0.5100	0.0681	0.9520	***
High - Control	0.5397	0.0977	0.9816	***
Low - VeryHigh	-0.6663	-1.1083	-0.2244	***
Low - Medium	-0.5401	-0.9820	-0.0981	***
Low - High	-0.5100	-0.9520	-0.0681	***
Low - Control	0.0296	-0.4123	0.4716	
Control - VeryHigh	-0.6960	-1.1379	-0.2540	***
Control - Medium	-0.5697	-1.0116	-0.1277	***
Control - High	-0.5397	-0.9816	-0.0977	***
Control - Low	-0.0296	-0.4716	0.4123	

15:51 Tuesday, July 3, 2012

The GLM Procedure

t Tests (LSD) for y

NOTE: This test controls the Type I comparisonwise error rate, not the experimentwise error rate.

Alpha	0.05
Error Degrees of Freedom	20
Error Mean Square	0.112222
Critical Value of t	2.08596
Least Significant Difference	0.442

Means with the same letter are not significantly different.

t Grouping	Mean	N	treat
A	1.3235	5	VeryHigh
A			
A	1.1972	5	Medium
A			
A	1.1672	5	High
B	0.6572	5	Low
B			
B	0.6275	5	Control

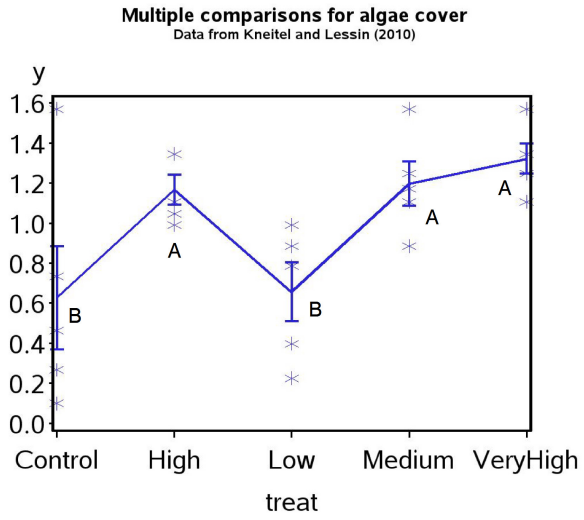

Figure 13.2: Algae cover vs. nutrient addition treatment for data from Kneitel and Lessin (2010). Means with different letters are significantly different (least significant difference method).

We will now calculate the value of LSD for this example to show how it is used to construct confidence intervals and tests. From the ANOVA output for `proc glm`, we see that $MS_{within} = 0.1122$ with 20 degrees of freedom. From Table T (Chapter 22), using $\alpha = 0.05$ we see that $c_{0.05,20} = 2.086$. There are also $n = 5$ replicates per treatment. We then have

$$LSD = c_{\alpha,a(n-1)}\sqrt{\frac{2MS_{within}}{n}} = 2.086\sqrt{\frac{2(0.1122)}{5}} = 0.4419. \tag{13.14}$$

Note that SAS also displays the value of LSD in the output. We next calculate a 95% confidence interval for $\mu_{\text{medium}} - \mu_{\text{control}}$. Recall that the formula for the interval is

$$\left(\bar{Y}_i - \bar{Y}_j - LSD, \bar{Y}_i - \bar{Y}_j + LSD\right). \tag{13.15}$$

Inserting the estimated means for these two treatments (see SAS output) in this formula, and the LSD value, we obtain

$$(1.1972 - 0.6275 - 0.4419, 1.1972 - 0.6275 + 0.4419) \tag{13.16}$$

or $(0.1278, 1.0116)$. This confidence interval and the LSD value are quite close to the values obtained by SAS.

We now show how the LSD value is used to test $H_0 : \mu_{\text{medium}} - \mu_{\text{control}} = 0$ or equivalently $H_0 : \mu_{\text{medium}} = \mu_{\text{control}}$. We would reject H_0 if $|\bar{Y}_i - \bar{Y}_j| \geq LSD$. Inserting the estimated means for these two treatments, we see that $|1.1972 - 0.6275| = 0.5687 \geq 0.4419$, and so this pair of means is significantly different.

13.3.3 The Tukey procedure

The Tukey method for multiple comparisons is similar to the least significant difference procedure, except that it uses the **studentized range distribution** in place of the t distribution. The studentized range distribution is designed to control the EER rate for all pairwise comparisons among group means (Hsu 1996). Another advantage is that the confidence intervals constructed using this distribution are **simultaneous confidence intervals**. This means that the overall probability the confidence intervals include the true value of $\mu_i - \mu_j$, for all pairs of groups, is equal to $1 - \alpha$ for some specified α. The overall probability α is also the EER for the family of all pairwise tests.

The Tukey procedure makes use of a quantity called the honestly significant difference (HSD), defined as

$$HSD = q_{\alpha,a,a(n-1)}\sqrt{\frac{MS_{within}}{n}}. \tag{13.17}$$

The quantity $q_{\alpha,a,a(n-1)}$ is obtained from the studentized range distribution, and depends on α (the desired EER), the number of groups a, as well as the degrees of freedom for MS_{within}.

To test $H_0 : \mu_i = \mu_j$ or $H_0 : \mu_i - \mu_j = 0$, we accept H_0 if $|\bar{Y}_i - \bar{Y}_j| < HSD$, and reject it $|\bar{Y}_i - \bar{Y}_j| \geq HSD$. This same rule applies to any pair of groups, because HSD would take the same value. Any pair of means that equals or exceeds this value is declared to be significantly different. The Tukey confidence intervals are of the form

$$\left(\bar{Y}_i - \bar{Y}_j - HSD, \bar{Y}_i - \bar{Y}_j + HSD\right). \tag{13.18}$$

13.3.4 Tukey procedure - SAS demo

Implementing the Tukey procedure requires only a small change in our previous SAS program. The `means` statement within `proc glm` becomes

```
means treat / tukey cldiff lines;
```

Confidence intervals for $\mu_i - \mu_j$ and $\mu_j - \mu_i$ are given for every pair of groups, as well as a diagram indicating which treatments are significantly different. See a section of the SAS output below. For this example, the Tukey finds fewer significant comparisons than the least significant difference procedure. We see there are only two significant comparisons, very high vs. low and very high vs. control treatments. This is a common pattern observed with multiple comparison tests, a few significant differences but also substantial overlap among treatments or groups.

SAS Output

Multiple comparisons for algae cover 4
Data from Kneitel and Lessin (2010)
11:45 Thursday, July 5, 2012

The GLM Procedure

Tukey's Studentized Range (HSD) Test for y

NOTE: This test controls the Type I experimentwise error rate.

Alpha	0.05
Error Degrees of Freedom	20
Error Mean Square	0.112222
Critical Value of Studentized Range	4.23186
Minimum Significant Difference	0.634

Comparisons significant at the 0.05 level are indicated by ***.

treat Comparison	Difference Between Means	Simultaneous 95% Confidence Limits		
VeryHigh - Medium	0.1263	-0.5077	0.7603	
VeryHigh - High	0.1563	-0.4777	0.7903	
VeryHigh - Low	0.6663	0.0323	1.3003	***
VeryHigh - Control	0.6960	0.0620	1.3300	***
Medium - VeryHigh	-0.1263	-0.7603	0.5077	
Medium - High	0.0300	-0.6040	0.6640	
Medium - Low	0.5401	-0.0939	1.1741	
Medium - Control	0.5697	-0.0643	1.2037	
High - VeryHigh	-0.1563	-0.7903	0.4777	
High - Medium	-0.0300	-0.6640	0.6040	
High - Low	0.5100	-0.1239	1.1440	
High - Control	0.5397	-0.0943	1.1737	
Low - VeryHigh	-0.6663	-1.3003	-0.0323	***
Low - Medium	-0.5401	-1.1741	0.0939	
Low - High	-0.5100	-1.1440	0.1239	
Low - Control	0.0296	-0.6044	0.6636	
Control - VeryHigh	-0.6960	-1.3300	-0.0620	***
Control - Medium	-0.5697	-1.2037	0.0643	

```
Control  - High          -0.5397    -1.1737  0.0943
Control  - Low           -0.0296    -0.6636  0.6044
```

```
                Multiple comparisons for algae cover                    5
                Data from Kneitel and Lessin (2010)
                                        11:45 Thursday, July 5, 2012

                          The GLM Procedure

                Tukey's Studentized Range (HSD) Test for y

NOTE: This test controls the Type I experimentwise error rate, but it generally
              has a higher Type II error rate than REGWQ.


              Alpha                                   0.05
              Error Degrees of Freedom                  20
              Error Mean Square                   0.112222
              Critical Value of Studentized Range  4.23186
              Minimum Significant Difference         0.634


        Means with the same letter are not significantly different.

  Tukey Grouping          Mean      N    treat

               A        1.3235      5    VeryHigh
               A
          B    A        1.1972      5    Medium
          B    A
          B    A        1.1672      5    High
          B
          B             0.6572      5    Low
          B
          B             0.6275      5    Control
```

We will now calculate the value of HSD for this example, to show how it is used to construct confidence intervals and tests. As before, we have $MS_{within} = 0.1122$ with 20 degrees of freedom. The SAS output gives the value of $q_{0.05,5,20} = 4.2319$, and there are $n = 5$ replicates per treatment. We then have

$$HSD = q_{\alpha,a,a(n-1)}\sqrt{\frac{MS_{within}}{n}} = 4.2319\sqrt{\frac{(0.1122)}{5}} = 0.6339. \qquad (13.19)$$

This value agrees with the SAS output labeled `Minimum Significant Difference`. We now calculate a 95% confidence interval for $\mu_{\text{medium}} - \mu_{\text{control}}$. The formula for the confidence interval is

$$\left(\bar{Y}_i - \bar{Y}_j - HSD, \bar{Y}_i - \bar{Y}_j + HSD\right). \qquad (13.20)$$

Inserting the estimated means for these two treatments (see SAS output) in this formula, and the HSD value, we obtain

$$(1.1972 - 0.6275 - 0.6339, 1.1972 - 0.6275 + 0.6339). \qquad (13.21)$$

or $(-0.0642, 1.2036)$. This confidence interval is close to the value provided by SAS.

How does this procedure control the EER as well as provide simultaneous confidence intervals? **The Tukey procedure basically controls the EER by making each pairwise test more conservative, through the use of the studentized range distribution.** Notice that $HSD > LSD$ for the same data set (0.6339 vs. 0.4419). This means that the Tukey procedure requires a larger difference between groups before declaring they are significantly different, and the confidence intervals are also broader. As a consequence, there is lower power to detect differences among groups when they do exist. This is the price paid for controlling the EER.

13.3.5 Multiple range tests - REGW

The multiple comparison procedures we have examined so far yield both tests and confidence intervals. Another type of multiple comparison procedure are multiple range tests. These procedures provide only tests, but are also more powerful procedures because they essentially conduct fewer overall tests than the methods we studied earlier. There are a number of

different multiple range tests, but we will only examine the REGW (Ryan-Einot-Gabriel-Welsch) procedure because it controls the EER (Hsu 1996).

The test works as follows (Hsu 1996). Suppose we order the sample means of the a different groups from smallest to largest:

$$\bar{Y}_{[1]} \leq \bar{Y}_{[2]} \leq \ldots \bar{Y}_{[a-1]}, \leq \bar{Y}_{[a]} \tag{13.22}$$

where $\bar{Y}_{[1]}$ is the smallest and $\bar{Y}_{[a]}$ the largest sample mean.

We then examine the range (difference) between the largest and smallest sample mean, namely $\bar{Y}_{[a]} - \bar{Y}_{[1]}$. If

$$\bar{Y}_{[a]} - \bar{Y}_{[1]} < q_a \sqrt{\frac{MS_{within}}{n}} \tag{13.23}$$

then we stop and declare there are no significant differences among groups. Otherwise, we assert that these two groups are significantly different and continue the process. We next examine the next innermost ranges $\bar{Y}_{[a-1]} - \bar{Y}_{[1]}$ and $\bar{Y}_{[a]} - \bar{Y}_{[2]}$. If

$$\bar{Y}_{[a-1]} - \bar{Y}_{[1]]} < q_{a-1} \sqrt{\frac{MS_{within}}{n}} \tag{13.24}$$

and

$$\bar{Y}_{[a1]} - \bar{Y}_{[2]]} < q_{a-1} \sqrt{\frac{MS_{within}}{n}} \tag{13.25}$$

then we stop the testing process. Otherwise, we assert that one or both groups are significantly different. This process is continued until no more significant differences are found.

The values of q are not the same for every step of the test. They are constructed so that $q_a > q_{a-1} > \ldots > q_2$, meaning that the largest range is tested using the largest value of q, the next largest two ranges with a smaller value of q, and so forth. This implies that the largest range must have the largest difference in means to be judged significant, while later tests allow for smaller differences. The values of q are chosen so that the experimentwise error rate has a specified value, usually $\alpha = 0.05$ (Hsu 1996). The studentized range distribution is involved in this process. The value of q_a used in the first step of the procedure is the same as that used by the Tukey procedure, as well as the difference in the means judged to be significant. The two procedures diverge after this point.

13.3.6 REGW procedure - SAS demo

Implementing the REGW procedure requires only a small change in our previous SAS programs. The `means` statement within `proc glm` becomes

```
means treat / regwq;
```

Here the `regwq` option requests the REGW procedure. SAS then generates a diagram indicating which groups are significantly different. See a section of the SAS output below, using the same data as our previous examples. For this example, the REGW procedure gives the same pattern of significant differences among groups as the Tukey method. The REGW procedure may become liberal (not fully control the EER) when the data are unbalanced, and SAS prints a warning note in this situation.

SAS Output

```
                    Multiple comparisons for algae cover                    4
                    Data from Kneitel and Lessin (2010)
                                          11:45 Thursday, July 5, 2012

                              The GLM Procedure

                Ryan-Einot-Gabriel-Welsch Multiple Range Test for y

              NOTE: This test controls the Type I experimentwise error rate.


                               Alpha                          0.05
                               Error Degrees of Freedom         20
                               Error Mean Square          0.112222


Number of Means              2             3             4             5
Critical Range       0.5340892     0.5871678     0.5930101     0.6339938

              Means with the same letter are not significantly different.

  REGWQ Grouping          Mean      N    treat

                A        1.3235      5    VeryHigh
                A
            B   A        1.1972      5    Medium
            B   A
            B   A        1.1672      5    High
            B
            B            0.6572      5    Low
            B
            B            0.6275      5    Control
```

13.4 Comparisons with a control - Dunnett procedure

Many studies include some sort of control group or treatment, and the experimenter may only be interested in comparing the control group with each of the other $a-1$ groups. For example, the control could represent a standard medical treatment for a disease while the other treatments represent alternative forms of therapy. The physician only wants to know if the alternative forms are better or worse than the standard method.

In this situation, there are only $a-1$ comparisons to be made rather than the full $a(a-1)/2$ comparisons of all pairs of means. The Dunnett procedure is designed to control the EER for just these $a-1$ comparisons, and hence has more power than other pairwise methods (Hsu 1996). The calculations are similar to the Tukey method, but use the quantity

$$DSD = d_{\alpha,a,a(n-1)}\sqrt{\frac{2MS_{within}}{n}}, \tag{13.26}$$

where DSD stands for Dunnett's significant difference. The values of $d_{\alpha,a,a(n-1)}$ are obtained from a distribution analogous to the studentized range distribution, except that it controls the EER for $a-1$ comparisons. The value of d depends on α (the desired EER), the number of groups a, and the degrees of freedom for MS_{within}.

Let μ_c be the mean of the control group, while μ_i is any other group. Dunnett's procedure can be used to test for $H_0 : \mu_i = \mu_c$ or equivalently $H_0 : \mu_i - \mu_c = 0$. We would accept H_0 if $|\bar{Y}_i - \bar{Y}_c| < DSD$. Conversely, we would reject H_0 if $|\bar{Y}_i - \bar{Y}_c| \geq DSD$. This same rule applies to all comparisons with the control group.

Confidence intervals for $\mu_i - \mu_c$ have the form

$$\left(\bar{Y}_i - \bar{Y}_c - DSD, \bar{Y}_i - \bar{Y}_c + DSD\right). \tag{13.27}$$

13.4.1 Dunnett's procedure - SAS demo

Using Dunnett's procedure requires only a small change to our program. The `means` statement within `proc glm` becomes

```
means treat / dunnett('Control');
```

The control group in our data set is coded as `Control`, and the `('Control')` portion of the statement informs SAS of this fact. Confidence intervals for $\mu_i - \mu_c$ are given in the SAS output, with the symbol *** indicating which comparisons of the control are significantly different. We see that the very high and medium treatments are significantly different from control.

SAS Output

```
                    Multiple comparisons for algae cover                    4
                    Data from Kneitel and Lessin (2010)
                                        11:45 Thursday, July 5, 2012

                              The GLM Procedure

                           Dunnett's t Tests for y

NOTE: This test controls the Type I experimentwise error for comparisons of all
                        treatments against a control.


                    Alpha                             0.05
                    Error Degrees of Freedom            20
                    Error Mean Square             0.112222
                    Critical Value of Dunnett's t  2.65103
                    Minimum Significant Difference  0.5617


          Comparisons significant at the 0.05 level are indicated by ***.


                                  Difference      Simultaneous
                 treat               Between     95% Confidence
              Comparison             Means         Limits

            VeryHigh - Control        0.6960     0.1343  1.2576  ***
            Medium   - Control        0.5697     0.0080  1.1314  ***
            High     - Control        0.5397    -0.0220  1.1013
            Low      - Control        0.0296    -0.5320  0.5913
```

13.5 Bonferroni and Sidak corrections

One way of controlling the EER in a set of comparisons is to use a distribution designed to control it, such as the studentized range distribution. These procedures control the EER by essentially making the per comparison rate for each test more conservative. This adjustment of the per comparison error rate is built into the studentized range distribution.

The Bonferroni correction provides another way of controlling the EER, by explicitly reducing the per comparison error rate and then using a simple t test (like the least significant difference procedure) to compare group means. Suppose that we are interested in k possible comparisons, either all $a(a-1)/2$ pairwise comparisons or $a-1$ comparisons with a control, where a is the number of groups. The Bonferroni correction adjusts the per comparison error rate as follows. Let α be the per comparison error rate, while α' is the desired EER. If we conduct each comparison at the per comparison rate of

$$\alpha = \frac{\alpha'}{k}, \tag{13.28}$$

then it can be shown the EER will not exceed α' (Hsu 1996). For example, suppose we are interested in all $k = a(a-1)/2$ pairwise comparison among groups. We would then conduct each test at the

$$\alpha = \frac{\alpha'}{k} = \frac{\alpha'}{a(a-1)/2} \tag{13.29}$$

level. We would use the same t test as in the least significant difference procedure, but adjust the value α according to this formula. We then have

$$BSD = c_{\frac{\alpha'}{a(a-1)/2}, a(n-1)} \sqrt{\frac{2MS_{within}}{n}}, \tag{13.30}$$

where BSD is the difference judged to be significant given the Bonferroni correction. We would accept $H_0 : \mu_i = \mu_j$ (or $H_0 : \mu_i - \mu_j = 0$) if $\bar{Y}_i - \bar{Y}_j$ falls inside the interval $(-BSD, BSD)$, or equivalently if $|\bar{Y}_i - \bar{Y}_j| < BSD$. Conversely, we would reject H_0 if $|\bar{Y}_i - \bar{Y}_j| \geq BSD$. A confidence interval for $\mu_i - \mu_j$ based on the Bonferroni correction would have the form

$$\left(\bar{Y}_i - \bar{Y}_j - BSD, \bar{Y}_i - \bar{Y}_j + BSD\right). \tag{13.31}$$

To make things more concrete, we can calculate the value of BSD for the algae cover example (Kneitel & Lessin 2010). From our previous output, we

have $a = 5$ groups, $n = 5$ replicates per group, and $MS_{within} = 0.1122$. If we set the EER to be $\alpha' = 0.05$, by the above formula we have

$$\alpha = \frac{\alpha'}{a(a-1)/2} = \frac{0.05}{5(5-1)/2} = \frac{0.05}{10} = 0.005. \tag{13.32}$$

For $\alpha = 0.005$, we have $c_{0.005,20} = 3.1534$, and so

$$BSD = c_{\frac{\alpha'}{a(a-1)/2},a(n-1)}\sqrt{\frac{2MS_{within}}{n}} = 3.1534\sqrt{\frac{2(0.1122)}{5}} = 0.6681. \tag{13.33}$$

Note that the value of $BSD = 0.6681$ is larger than $HSD = 0.6339$ value for the Tukey procedure. Thus, the Bonferroni method requires a greater difference among means before declaring they are significantly different, implying it has lower power than the Tukey procedure. It would also generate larger confidence intervals and so provides less precision in estimation.

Given these drawbacks, why would the Bonferroni correction be used? The Bonferroni procedure is quite general and can be used to control the EER for other testing procedures, not just comparisons among means in ANOVA. For example, it is common to have a collection of statistical tests that address a particular question. We might have a single experiment in which a number of different Y variables are measured, with a separate ANOVA conducted on each variable. If enough variables are examined it is possible that some could be significant by chance, and we could control the EER for all these tests using the Bonferroni correction, with k being the number of Y variables. There is also a version of this procedure similar in spirit to REGW, called the **sequential Bonferroni method** (Rice 1989). The sequential Bonferroni alleviates to some extent the lack of power in the standard Bonferroni correction. This procedure is implemented in `proc multtest` in SAS.

The Sidak correction is another procedure used to control the EER, which provides slightly more power than the Bonferroni method. Let α be the per comparison error rate, while α' is the desired EER. If we conduct each comparison at the per comparison rate of

$$\alpha = 1 - (1 - \alpha')^{1/k}, \tag{13.34}$$

then the actual EER will not exceed α'. For example, suppose we are interested in all $k = a(a-1)/2$ pairwise comparison among groups. We would then conduct each test at the

$$\alpha = 1 - (1 - \alpha')^{1/k} = 1 - (1 - \alpha')^{1/[a(a-1)/2]} \tag{13.35}$$

level. For $\alpha' = 0.05$ and $a = 5$ groups, we obtain

$$\alpha = 1 - (1 - \alpha')^{1/[a(a-1)/2]} = 1 - (1 - 0.05)^{1/10} = 0.0051. \quad (13.36)$$

We would then compare pairs of means using the same test as for the Bonferroni correction, except that we would use $\alpha = 0.0051$ rather than $\alpha = 0.005$. This value of α is a bit larger than the corresponding Bonferroni one, making the Sidak correction slightly more powerful.

SAS implements both the Bonferroni and Sidak corrections in the `means` statement with the options `bon` or `sidak`, similar to using the `tukey` option.

13.6 Vascular plant cover - SAS demo

Kneitel & Lessin (2010) also examined vascular plant cover in their study of the effect of eutrophication on vernal pools in California. Vascular plant cover (`cover`) was derived by subtracting algal cover (`algae`) from total cover (`total`), then arcsine-square root transformed before analysis (see Chapter 15). See `data` step in the SAS program below.

The `proc glm` code compares all possible pairs of group means using the Tukey procedure, and also compares the `Control` treatment with the other treaments using Dunnett's procedure. This was done to provide more examples of these procedures. **In practice, you should choose one procedure for comparing the means.**

The diagram generated by the Tukey procedure indicates two significant differences among treatments. Reading the diagram, we see the control vs. high and control vs. very high comparisons are significant, because they have different letters. No other pairs of groups are significantly different. Fig. 13.3 indicates how these results could be graphically displayed using letters. We see that vascular plant cover actually decreases with increased nutrient levels, likely due to inhibition from the algal mats that form (Kneitel and Lessin 2010).

We can also determine which groups are significantly different by examining the confidence intervals generated by the Tukey procedure. Confidence intervals that do not include zero indicate a significant difference among groups, because of the duality between confidence intervals and tests (see Chapter 10). The significant tests are indicated by *** in the SAS output. The SAS output for Dunnett's procedure shows that the high and very high treatments are significantly different from the control group.

SAS Program

```
* Kneitel_2010_cover2.sas;
options pageno=1 linesize=80;
goptions reset=all;
title 'Multiple comparisons for vascular plant cover';
title2 'Data from Kneitel and Lessin (2010)';
data kneitel;
        input treat $ richness total algae;
        * Apply transformations here;
        vcover = total-algae;
        y = arsin(sqrt(vcover/100));
        datalines;
Control   8   78    1
Control   5   84    7
Control  10         115   45
Control   7         200  100
Control   6   72   20
Low       8   73   15
Low       7  124   70
Low       8  116   50
Low       8   92    5
Low       7  138   60
Medium    7  124   85
Medium    8  116   80
Medium    8  145   60
Medium    6  154  100
Medium    7  129   90
High      6  134   95
High      7  138   95
High      8  103   70
High      8  119   75
High          6  132   80
VeryHigh  6  148   95
VeryHigh  5  134   95
VeryHigh  5  119  100
VeryHigh  5  117   90
VeryHigh  5  129   80
;
run;
* Print data set;
proc print data=kneitel;
* Plot means, standard errors, and observations;
proc gplot data=kneitel;
        plot y*treat=1 / vaxis=axis1 haxis=axis1;
```

```
        symbol1 i=std1mjt v=star height=2 width=3;
        axis1 label=(height=2) value=(height=2) width=3 major=(width=2) minor=none;
run;
* One-way anova with comparisons;
proc glm order=data data=kneitel;
        class treat;
        model y = treat;
        output out=resids p=pred r=resid;
        * Tukey procedure - controls the EER;
        means treat / tukey cldiff lines;
        * Dunnett's procedure - controls EER for comparisons with a control;
        means treat / dunnett('Control');
run;
goptions reset=all;
title "Diagnostic plots to check anova assumptions";
* Plot residuals vs. predicted values;
proc gplot data=resids;
        plot resid*pred=1 / vaxis=axis1 haxis=axis1;
        symbol1 v=star height=2 width=3;
        axis1 label=(height=2) value=(height=2) width=3 major=(width=2) minor=none;
run;
* Normal quantile plot of residuals;
proc univariate noprint data=resids;
        qqplot resid / normal waxis=3 height=4;
run;
quit;
```

SAS Output

Multiple comparisons for vascular plant cover 1
Data from Kneitel and Lessin (2010)
11:45 Thursday, July 5, 2012

Obs	treat	richness	total	algae	vcover	y
1	Control	8	78	1	77	1.07062
2	Control	5	84	7	77	1.07062
3	Control	10	115	45	70	0.99116
4	Control	7	200	100	100	1.57080
5	Control	6	72	20	52	0.80540
6	Low	8	73	15	58	0.86574
7	Low	7	124	70	54	0.82544
8	Low	8	116	50	66	0.94826
9	Low	8	92	5	87	1.20193
10	Low	7	138	60	78	1.08259
11	Medium	7	124	85	39	0.67449
12	Medium	8	116	80	36	0.64350
13	Medium	8	145	60	85	1.17310
14	Medium	6	154	100	54	0.82544
15	Medium	7	129	90	39	0.67449
16	High	6	134	95	39	0.67449
17	High	7	138	95	43	0.71517
18	High	8	103	70	33	0.61194
19	High	8	119	75	44	0.72525
20	High	6	132	80	52	0.80540
21	VeryHigh	6	148	95	53	0.81542
22	VeryHigh	5	134	95	39	0.67449
23	VeryHigh	5	119	100	19	0.45103
24	VeryHigh	5	117	90	27	0.54640
25	VeryHigh	5	129	80	49	0.77540

Multiple comparisons for vascular plant cover 2
Data from Kneitel and Lessin (2010)
11:45 Thursday, July 5, 2012

The GLM Procedure

Class Level Information

Class	Levels	Values

```
          treat             5    Control Low Medium High VeryHigh


                   Number of Observations Read          25
                   Number of Observations Used          25


                Multiple comparisons for vascular plant cover                3
                     Data from Kneitel and Lessin (2010)
                                              11:45 Thursday, July 5, 2012

                              The GLM Procedure

Dependent Variable: y

                                      Sum of
 Source                       DF       Squares    Mean Square   F Value   Pr > F

 Model                         4    0.71900305     0.17975076      4.93   0.0063

 Error                        20    0.72959178     0.03647959

 Corrected Total              24    1.44859482


              R-Square     Coeff Var      Root MSE        y Mean

              0.496345      22.50344      0.190996      0.848743


 Source                       DF      Type I SS    Mean Square   F Value   Pr > F

 treat                         4    0.71900305     0.17975076      4.93   0.0063


 Source                       DF    Type III SS    Mean Square   F Value   Pr > F

 treat                         4    0.71900305     0.17975076      4.93   0.0063


                Multiple comparisons for vascular plant cover                4
                     Data from Kneitel and Lessin (2010)
                                              11:45 Thursday, July 5, 2012

                              The GLM Procedure
```

Tukey's Studentized Range (HSD) Test for y

NOTE: This test controls the Type I experimentwise error rate.

Alpha	0.05
Error Degrees of Freedom	20
Error Mean Square	0.03648
Critical Value of Studentized Range	4.23186
Minimum Significant Difference	0.3615

Comparisons significant at the 0.05 level are indicated by ***.

treat Comparison	Difference Between Means	Simultaneous 95% Confidence Limits		
Control - Low	0.1169	-0.2445	0.4784	
Control - Medium	0.3035	-0.0580	0.6650	
Control - High	0.3953	0.0338	0.7567	***
Control - VeryHigh	0.4492	0.0877	0.8106	***
Low - Control	-0.1169	-0.4784	0.2445	
Low - Medium	0.1866	-0.1749	0.5481	
Low - High	0.2783	-0.0831	0.6398	
Low - VeryHigh	0.3322	-0.0292	0.6937	
Medium - Control	-0.3035	-0.6650	0.0580	
Medium - Low	-0.1866	-0.5481	0.1749	
Medium - High	0.0918	-0.2697	0.4532	
Medium - VeryHigh	0.1457	-0.2158	0.5071	
High - Control	-0.3953	-0.7567	-0.0338	***
High - Low	-0.2783	-0.6398	0.0831	
High - Medium	-0.0918	-0.4532	0.2697	
High - VeryHigh	0.0539	-0.3076	0.4154	
VeryHigh - Control	-0.4492	-0.8106	-0.0877	***
VeryHigh - Low	-0.3322	-0.6937	0.0292	
VeryHigh - Medium	-0.1457	-0.5071	0.2158	
VeryHigh - High	-0.0539	-0.4154	0.3076	

The GLM Procedure

Tukey's Studentized Range (HSD) Test for y

NOTE: This test controls the Type I experimentwise error rate, but it generally has a higher Type II error rate than REGWQ.

Alpha	0.05
Error Degrees of Freedom	20
Error Mean Square	0.03648
Critical Value of Studentized Range	4.23186
Minimum Significant Difference	0.3615

Means with the same letter are not significantly different.

Tukey Grouping	Mean	N	treat
A	1.1017	5	Control
A			
B A	0.9848	5	Low
B A			
B A	0.7982	5	Medium
B			
B	0.7065	5	High
B			
B	0.6525	5	VeryHigh

Multiple comparisons for vascular plant cover 6
Data from Kneitel and Lessin (2010)
11:45 Thursday, July 5, 2012

The GLM Procedure

Dunnett's t Tests for y

NOTE: This test controls the Type I experimentwise error for comparisons of all treatments against a control.

Alpha	0.05

```
Error Degrees of Freedom            20
Error Mean Square                0.03648
Critical Value of Dunnett's t    2.65103
Minimum Significant Difference    0.3202

Comparisons significant at the 0.05 level are indicated by ***.

                          Difference      Simultaneous
          treat             Between   95% Confidence
       Comparison             Means        Limits

Low      - Control          -0.1169    -0.4372  0.2033
Medium   - Control          -0.3035    -0.6237  0.0167
High     - Control          -0.3953    -0.7155 -0.0750  ***
VeryHigh - Control          -0.4492    -0.7694 -0.1289  ***
```

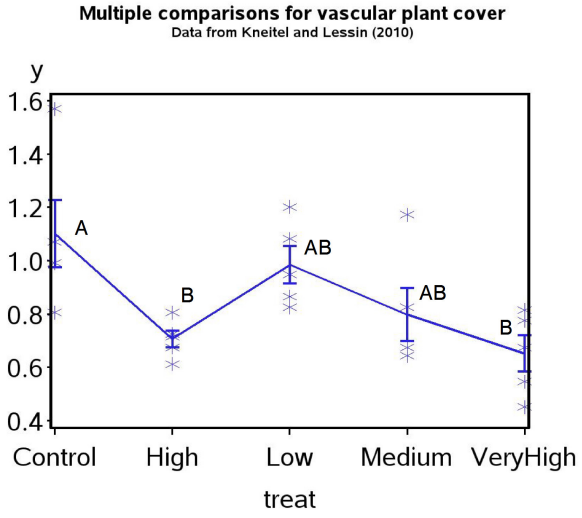

Figure 13.3: Vascular plant cover vs. nutrient addition treatment for simulated data patterned after Kneitel and Lessin (2010). Means with different letters are significantly different (Tukey procedure).

13.7 False discovery rate method

The multiple comparison procedures we have examined control the EER, but at the cost of power. This is especially true for studies with many treatments or groups. For example, suppose we have $a = 5$ treatments and want to conduct all pairwise comparisons using the Bonferroni method, with an EER of $\alpha' = 0.05$. There are $k = a(a-1)/2 = 5(4)/2 = 10$ pairwise comparisons, and so we would conduct each comparison at the $\alpha = \alpha'/k = 0.05/10 = 0.005$ level. For $a = 10$ treatments, a similar calculation suggests that each comparison should be conducted at the $\alpha = 0.0011$ level, yielding a much more conservative test. As the number of treatments increases, this makes it less likely significant differences will be found, and so the power to detect differences among treatments decreases. The number of treatments has similar effects on other multiple comparison procedures that control the EER.

The **false discovery rate** method provides an alternative approach to multiple comparisons and tests. This method controls the **proportion** of Type I errors in a set of comparisons, known as the false discovery rate or FDR (Benjamini & Hochberg 1995). This differs substantially from methods that control the EER, which are concerned with keeping the **number** of Type I errors low. One will have more Type I errors using the FDR, but the proportion of them is controlled, and the power to detect differences among treatments will be higher than EER methods. This approach seems particularly useful for studies that screen many treatments or groups, possibly for future work, and it is more important to identify possible effects than controlling the number of Type I errors (Verhoeven et al. 2005).

The FDR method for multiple comparisons works as as follows (Benjamini & Hochberg 1995). Suppose you have k pairwise comparisons, and obtain a P value for each one using the LSD procedure. Let $P_{[1]} \leq P_{[2]} \leq \ldots \leq P_{[k]}$ be the P values for these tests, ordered from smallest to largest, with $P_{[i]}$ the ith one. Let α^* be the specified false discovery rate. We then examine the ordered P values from largest to smallest (from $i = k$ to 1), examining at each step whether

$$P_{[i]} \leq \frac{i}{k}\alpha^*. \tag{13.37}$$

We can see that the right side of this equation decreases from α^* to α^*/k as i decreases. The first time this inequality is true, we declare that this pairwise comparison and all further ones are significantly different. Benjamini &

Hochberg (1995) show that this procedure controls the false discovery rate. The same method can also be used in other multiple testing scenarios, not just multiple comparisons among means.

As an example of this procedure, consider the algae cover example we examined earlier (Kneitel and Lessin 2010). There are ten pairwise comparisons among the different nutrient treatments. We first obtain the P values for each comparison using the LSD method (see SAS demo below), and order these from largest to smallest (Table 13.1). We then compare the P values with the right side of Eq. 13.37, beginning at the top of the table. We see that first comparison that satisfies Eq. 13.37 is high vs. low, and so we declare this comparison and all further ones to be significant. Thus, the the FDR procedure finds six of ten pairwise comparisons to be significant, similar to the LSD procedure. The Tukey and REGW procedures, which control the EER, found only two significant comparisons.

Table 13.1: Ordered P values for LSD comparisons of algae cover in different nutrient treatments (Kneitel and Lessin 2010). The last column calculates the right side of Eq. 13.37 for $\alpha^* = 0.05$ and $k = 10$ pairwise comparisons.

Comparison	i	$P_{[i]}$	$\frac{i}{k}\alpha^*$
control–low	10	0.8902	0.0500
medium–high	9	0.8887	0.0450
medium–very high	8	0.5578	0.0400
high–very high	7	0.4693	0.0350
high–low	6	0.0258	0.0300
control–high	5	0.0192	0.0250
low–medium	4	0.0191	0.0200
control–medium	3	0.0141	0.0150
low–very high	2	0.0051	0.0100
control–very high	1	0.0037	0.0010

13.7.1 False discovery rate - SAS demo

The FDR procedure can be implemented in two steps using SAS. We first need to obtain the P values for the LSD procedure. This can be accomplished by adding an `lsmeans` statement to our previous program, with a `pdiff` option:

```
lsmeans treat / adjust=t pdiff;
```

The result is a table of P values for each comparison, shown below.

SAS Output

```
                    Multiple comparisons for algae cover                    4
                    Data from Kneitel and Lessin (2010)
                                          14:39 Monday, May 23, 2016

                            The GLM Procedure
                           Least Squares Means

                                                    LSMEAN
                    treat            y LSMEAN       Number

                    Control        0.62753783            1
                    High           1.16721374            2
                    Low            0.65716894            3
                    Medium         1.19723297            4
                    VeryHigh       1.32351133            5


                    Least Squares Means for effect treat
                    Pr > |t| for H0: LSMean(i)=LSMean(j)

                           Dependent Variable: y

  i/j             1             2             3             4             5

    1                       0.0192        0.8902        0.0141        0.0037
    2        0.0192                       0.0258        0.8887        0.4693
    3        0.8902        0.0258                       0.0191        0.0051
    4        0.0141        0.8887        0.0191                      0.5578
    5        0.0037        0.4693        0.0051        0.5578
```

We then use `proc multtest` to carry out the FDR procedure. The P values for each comparison are supplied in a SAS data set, labeled as `raw_p`. The data set is specified using the `inpvalues` option, while the FDR procedure is requested using the `fdr` option. The output consists of the original and adjusted P values, with the adjustment made according to the FDR procedure. Adjusted P values less than 0.05 are judged to be significant. See program and output below. We observe that six of ten pairwise comparisons have an adjusted P value less than 0.05, and so these are judged significant by the FDR procedure.

SAS Program

```
* Kneitel_2010_algae_fdr2.sas;
options pageno=1 linesize=80;
goptions reset=all;
title 'Multiple comparisons for algae cover';
title2 'False discovery rate (Benjamini and Hochberg 1995)';
data pvalues;
        input comparison :$18. raw_p;
        datalines;
Control-High     0.0192
Control-Low      0.8902
Control-Medium   0.0141
Control-VeryHigh 0.0037
High-Low         0.0258
High-Medium      0.8887
High-VeryHigh    0.4693
Low-Medium       0.0191
Low-VeryHigh     0.0051
Medium-VeryHigh  0.5578
;
* Multiple comparisons using fdr;
proc multtest inpvalues=pvalues fdr;
run;
quit;
```

SAS Output

Multiple comparisons for algae cover 1
False discovery rate (Benjamini and Hochberg 1995)
14:39 Monday, May 23, 2016

The Multtest Procedure

P-Value Adjustment Information

P-Value Adjustment False Discovery Rate

p-Values

Test	Raw	False Discovery Rate
1	0.0192	0.0384
2	0.8902	0.8902
3	0.0141	0.0384
4	0.0037	0.0255
5	0.0258	0.0430
6	0.8887	0.8902
7	0.4693	0.6704
8	0.0191	0.0384
9	0.0051	0.0255
10	0.5578	0.6973

13.8 References

Benjamini, Y. & Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)* 57: 289-300.

Day, R. W. & Quinn, G. P. (1989) Comparisons of treatments after an analysis of variance in ecology. *Ecological Monographs* 59: 433-463.

Hsu, J. C. (1996) *Multiple Comparisons: Theory and Methods.* Chapman & Hall/CRC Press, Boca Raton, FL.

Kneitel, J. M. & Lessin, C. L. (2010) Ecosystem-phase interactions: aquatic eutrophication decreases terrestrial plant diversity in California vernal pools. *Oecologia* 163: 461-469.

Kohler, C. K, Heidinger, R. C. & Call, T. (1990) Levels of PCBs and trace metal in Crab Orchard Lake sediment, benthos, zooplankton and fish. Waste Management and Research Center Report RR-E43, Illinois Department of Natural Resources.

Rice, W. R. (1989) Analyzing tables of statistical tests. *Evolution* 43: 223-225.

SAS Institute Inc. (2014a) *SAS/STAT 13.2 Users Guide.* SAS Institute Inc., Cary, NC.

Verhoeven, K. J. F., Simonsen, K. L. & McIntyre, L. M. (2005) Implementing false discovery rate control: increasing your power. *Oikos* 108: 643-647.

Westfall, P. H., Tobias, R. D., Rom, D., Wolfinger, R. D. & Hochberg, Y. (1999) *Multiple Comparisons and Multiple Tests Using the SAS System.* SAS Institute Inc., Cary, NC.

13.9 Problems

1. White-tailed deer are voracious consumers of landscaping plants. A frustrated homeowner/professor is interested in testing whether different repellents actually reduce deer herbivory. Replicate plots of houseplants are established and four different treatments applied to the plots: (1) a control with no treatment, (2) hot pepper oil repellent, (3) rotten egg repellent, and (4) livestock blood repellent. There were 4 replicate plots per treatment. The amount of herbivory (percentage of plants eaten) after one month are given in the following table.

Control	Hot pepper	Rotten eggs	Blood
61.1	54.4	32.0	36.2
64.9	67.9	28.5	38.3
61.6	54.6	21.6	31.1
67.8	58.1	38.8	44.1

   (a) Test whether there is an overall effect of treatment on the percentage of plants eaten, using one-way anova and SAS. Report your results using P values and discuss the significance of the test.

   (b) Use the Tukey procedure to compare the different treatments, and interpret your results. Which pairs of treatments are significantly different? Do the treatments fall into particular groups?

   (c) Suppose the homeowner is only interested in treatments that are different from the control. Use the Dunnett method to compare the three treatments with the control one. Which treatments are significantly different from the control?

2. PCB concentrations were measured in the sediment of Crab Orchard Lake, at 11 different sites (Kohler et al. 1990). Three samples were taken at each site, yielding the data shown in the table below. Site 10 is near an abandoned dump site for a manufacturer of electrical transformers.

Site	PCB (mg/kg), sample 1-3
1	0.0453, 0.0626, 0.527
2	0.0395, 0.0494, 0.0416
3	0.0234, 0.0451, 0.0541
4	0.033, 0.0643, 0.0517
5	0.0394, 0.0810, 0.0266
6	0.0294, 0.0425, 0.0538
7	0.0255, 0.0440, 0.0427
8	0.0323, 0.0382, 0.0360
9	0.0533, 0.0407, 0.0626
10	0.160, 0.437, 0.343
11	0.135, 0.142, 0.0592

(a) Test whether there is an overall effect of site on PCB concentration, using one-way ANOVA and SAS. Treat site as a fixed effect. Report your results using P values and discuss the significance of the test. A log transformation should be applied before analysis.

(b) Use the REGW procedure to compare the different sites, and interpret your results. Which pairs of sites are significantly different? Do the sites fall into particular groups?

3. An entomologist wants to compare the attractiveness of nine different baits (A-I) for bark beetles. There were three replicate traps for each bait treatment. The table below lists the number of beetles captured in each trap.

Bait	Beetles, trap 1-3
A	27, 36, 26
B	25, 19, 37
C	8, 16, 12
D	15, 8, 12
E	68, 42, 57
F	43, 32, 47
G	10, 12, 19
H	71, 62, 53
I	19, 11, 21

(a) Test whether there is an overall effect of bait on beetle captures, using one-way ANOVA and SAS. Report your results using P

values and discuss the significance of the test. Apply a log transformation before analysis.

(b) Use the FDR procedure to compare the different baits, and interpret your results. Which baits are significantly different?