Contrasts and Multiple Comparisons
Supplement for Section 6.5

Brian Habing – University of South Carolina
Last Updated: February 23, 2003

As discussed in Section 6.4 and at the beginning of Section 6.5, the *F*-test from the ANOVA table allows us to test the null hypothesis "The population means of all of the groups/treatments are equal." The alternate hypothesis is simply that "At least two are not equal." Often this isn't what we want to know!

Say we are comparing 20 possible treatments for a disease. The ANOVA *F*-test (sometimes called the omnibus test), could only tell us that at least one of the treatments worked differently than the others. We might, however, want to be able to rank the 20 from best to worst, and say which of these differences are significant. We might want to compare all the treatments produced by one company to those of another, or maybe all the treatments based on one idea to those based on another.

An obvious suggestion in each of these cases would be to simply do a large number of t-tests. To rank the 20 from best to worst, we could simply do a separate t-test for each possible comparison (there are 190 of them). To compare the two companies or two ideas, we could simply group all of the observations from the related methods together and use t-tests to see if they differ. One difficulty with this is that the α-level (probability of a Type I error) may no longer be what we want it to be.

Sidak's Formula

Stepping back from the ANOVA setting for a minute, say we wish to conduct one-sample t-tests on twenty completely independent populations. If we set $\alpha = 0.05$ for the first test, that means that:

$$0.05 = \alpha = P[\text{reject } H_0 \text{ for test one} \mid H_0 \text{ is true for test one}]$$

We could write the same for the other nineteen populations as well. If we are concerned about all twenty populations though, we might be more interested in the probability that we reject a true null hypothesis at all. That is,

$$\alpha_T = P[\text{reject } H_0 \text{ for test one} \cup \text{reject } H_0 \text{ for test two} \cup \cdots \cup \text{reject } H_0 \text{ for test 20} \mid H_0 \text{ is true for all tests}]$$

We call this quantity the *family-wise* (or *experiment-wise*) *error* rate. The α for each individual test is called the *comparison-wise* error rate. The family (or experiment), in this case, is made up of the twenty individual comparisons.

Using the rules of probability, and the fact that we assumed the tests were independent for this example, we can calculate what α_T would be if we used $\alpha = 0.05$ for the comparison-wise rate.

α_T = P[reject H$_0$ for 1 $\cup$ reject H$_0$ for 2 $\cup \cdots \cup$ reject H$_0$ for 20 | H$_0$ is true for all tests]

$\quad = 1 - $ P[fail to reject H$_0$ for 1 $\cap \cdots \cap$ fail to reject H$_0$ for 20| H$_0$ is true for all tests]

$\quad = 1 - $ P[fail to reject H$_0$ for 1| H$_0$ is true for all tests] $\cdots$ P[fail to reject H$_0$ for 2| H$_0$ is true for all tests]

$\quad = 1 - (1-\alpha)(1-\alpha) \cdots (1-\alpha) = 1 - (1-\alpha)^{20}$

$\quad = 1 - (1-0.05)^{20}$

$\quad = 1 - 0.95^{20}$

$\quad \approx 0.64$

The chance of making at least one error (α_T) isn't 5%, it's nearly 64%!

If we replace the twenty tests with k tests, we get *Sidak's formula*:

$$\alpha_T = 1 - (1-\alpha)^k$$

when the tests are independent. If we know what α_T we want, we can solve for the needed α, to get:

$$\alpha = 1 - (1-\alpha_T)^{1/k}$$

If we wanted $\alpha_T = 0.05$, this formula would show us that we need to use an α of 0.00256 for each individual comparison!
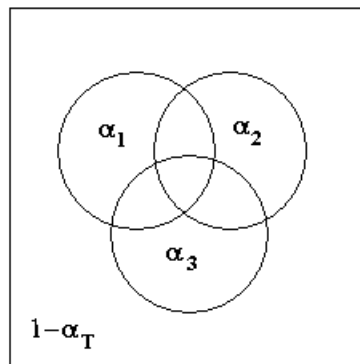

Bonferroni's Formula

In the case of ANOVA, the various tests will often not be independent. If we want to conduct the t-tests to compare 20 possible medical treatments to each other, then clearly the comparison of 1 to 2, and 1 to 3 will not be independent; they both contain treatment 1!

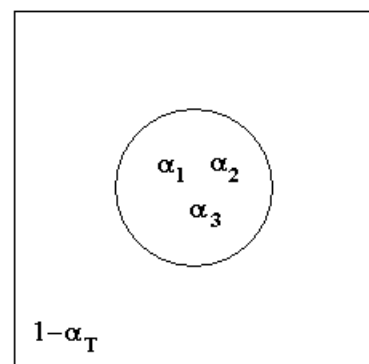The diagram below illustrates three possible situations that could occur for three tests:
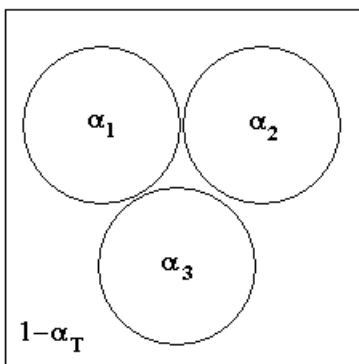
α_T is as large as possible	α_T is in between	α_T is as small as possible
assume the worst	what usually happens	assume the best
Bonferroni	????	Fisher

The worst possible case in terms of α_T would be if the type I errors for the individual tests were mutually exclusive. In this case,

α_T = P[reject H_0 for 1 $\cup$ reject H_0 for 2 $\cup \cdots \cup$ reject H_0 for k | H_0 is true for all tests]

   = P[reject H_0 for 1 | H_0 is true for all tests] $+ \cdots +$ P[reject H_0 for k | H_0 is true for all tests]

   $= \alpha + \alpha + \cdots \alpha = k\alpha$ to a maximum of one.

or equivalently $\alpha = \alpha_T/k$. This is *Bonferroni's formula*.

The best possible case in terms of α_T would be if the type I errors for the individual tests all overlapped. In this case, $\alpha_T = \alpha$.

So far then...

If we are performing a set of tests that are independent, then we can use Sidak's adjustment to figure out what comparison-wise α we should be using.

If the tests are not independent, then we have a choice. We could be *liberal* and reject true null hypotheses too often (use $\alpha_T = \alpha$) or be *conservative* and not reject the true null hypotheses as much as we should for our desired α_T (use Bonferroni). In terms of α_T, we would be better being conservative then. The problem with this is that if we do not reject the true null hypotheses enough, we also will not reject the false ones enough!

In the case of comparing the means of treatments, if we are liberal (using $\alpha_T = \alpha$) we will "find" lots of differences that are there, but also lots of differences that aren't real! If we are conservative we won't find lots of fake differences, but we will also miss the real ones.

Fisher's LSD

One method for dealing with the fact that using $\alpha_T = \alpha$ is too liberal is called the Fisher Least Significant Difference (LSD) test (pages 254-256). The idea is to only check to see if the means of groups are different if you reject the omnibus F-test. This makes some obvious sense, if you fail to reject that there are no differences, why would you continue looking? While this helps keep the number of false rejections down, it does have two downsides. The first problem can occur when you fail to reject the overall ANOVA null hypothesis. Because the omnibus test from the ANOVA table is looking at all of the groups at once, it will sometimes miss a difference between just two means. It has to sacrifice power for each individual comparison in order to test them all at once. The second problem can occur when we do reject the overall ANOVA null hypothesis and proceed to do the other comparisons of the group means. The omnibus test may have rejected because of a difference between only two means, but because using $\alpha_T = \alpha$ is liberal, we may find more differences than are really there. Because of these two difficulties, Fisher's LSD can't be highly recommended.

The Holm Test

The Holm test is a method for dealing with the fact that the Bonferroni procedure is too conservative. The main idea comes from noticing that we always used the condition "H_0 is true for all tests", instead of using the condition that it was true only for the specific test we were doing. The procedure behind the Holm test is to first find all of the p-values for all of the individual tests we were performing, and then rank them from smallest to largest. Compare the smallest to $\alpha = \alpha_T/k$. If you fail to reject the null hypothesis for the first step, then you stop here. If you do reject, then compare the next smallest to $\alpha = \alpha_T/(k-1)$. Again, if you fail to reject the null hypothesis then you stop here; if you do reject continue on and use $\alpha = \alpha_T/(k-2)$. (You do not need to check the omnibus F-test first, thus avoiding the first problem with Fisher's LSD.)

For example, say you have five hypotheses you are testing, you wanted $\alpha_T = 0.05$, and you observed p-values of 0.011, 0.751, 0.020, 0.030, and 0.001 respectively

```
Test Number        P-value        Compare To                    Conclusion
     5             0.001      0.05/5=0.01            reject H₀ for test 5
     1             0.011      0.05/4=0.0125          reject H₀ for test 1
     3             0.020      0.05/3=0.0166          fail to reject for test 3
     4             0.030      no comparison made     fail to reject for test 4
     2             0.751      no comparison made     fail to reject for test 2
```

Notice that Bonferonni's test would only have rejected for test 5. Using $\alpha_T = \alpha$ would have rejected for tests 5, 1, 3, and 4. Thus the power of the Holm test is somewhere in between that of the Bonferroni procedure and Fisher's LSD.

While it is more powerful than Bonferroni's method (it rejects more false H_0's) it still makes sure that α_T is held to the desired level (unlike Fisher's LSD). Notice that if all the null hypotheses are true, we make an error if we reject any of them. The chance that we reject any is the same as the chance that we reject the first, which is α_T/k. We are thus safe for the same reason that Bonferroni's formula works. Now assume that we rejected the first null hypothesis because it was false. There are only $k-1$ tests left, and so when we go to the second test we can start as if we were using Bonferroni's formula with $k-1$ instead of k. And we continue in this way. While this argument is not a proof that the Holm Test protects the family-wise error rate α_T, it should make the general idea fairly clear.

While there are many other methods for making multiple comparisons (see pages 256-267), the Holm test performs fairly well compared to all of them, controls α_T at the desired level, and is fairly easy to understand. Because of this, it will be the method that we will focus on.

Contrasts

In order to perform any of these tests though, we must be able to tell SAS what we want done. The building blocks for many of the SAS procedures that we will have SAS use are called *contrasts*. They are discussed on pages 243-249 of the text.

The key formulas are that for the contrast $L = \sum\limits_{i=1}^{t} a_i \mu_i$ (where $\sum\limits_{i=1}^{t} a_i = 0$) are

4

$$\hat{L} = \sum_{i=1}^{t} a_i \bar{y}_i \qquad \text{and} \qquad \hat{\sigma}_{\hat{L}} = \sqrt{MSW \sum_{i=1}^{t} \frac{a_i}{n_i}} = \sqrt{MSE \sum_{i=1}^{t} \frac{a_i}{n_i}}$$

where $\hat{L}$ is normally distributed if the ANOVA assumptions are met. Since we have the standard error for $\hat{L}$, we could make a confidence interval for L, or test the null hypothesis that $L=0$. This is exactly what we will ask SAS to do in the examples below.

Independence, Orthogonal Contrasts and Sidak's Formula (an aside-note)

As discussed on pages 246- 249, two contrasts are said to be orthogonal if the dot-product (sum of the cross products) of their coefficient vectors is zero. The reason to care if two contrasts are orthogonal is that the estimates that go with a set of orthogonal contrasts are independent.   This will play an important roll in constructing the ANOVA table for factorial designs in Chapter 9. In the present situation though the test statistics will not be independent, however, as they both contain the same MSE in the denominator. Because of this it is not technically appropriate to use Sidak's formula even in this situation.

Tying it All Together

When we approach an ANOVA problem, there are three basic types of questions we could have in mind.

1. **Are there any differences between any of the group means?**
   Choose α and simply use the F test from the ANOVA table (the omnibus test).

2. **Do the means of some particular groups differ from the means of some other particular groups?** Choose α_T and come up with the contrasts you wish to test. Find the p-values for the tests that go with these contrasts, and then use the Holm test procedure to see which are significant.

3. **What is the order of the group means, and which are significantly different from each other?**
   Choose α_T.   Make all of the contrasts that compare two means to each other, find their p-values, and use the Holm test procedure to see which are significantly different. Then make a simple graph to display the result.

It is important to note that you should decide which one of these questions you want to answer before you look at any of the output. (If for some reason you don't know why you are looking at the data in advance, tthen Scheffé's method, discussed on pages 259-260, can be used.) Also, you should only pick one of these three questions. (It doesn't make sense to look at more than one of them, does it?) Finally, in all cases remember to check the assumptions!

Example - Shrimp Larvae Diets

The follow pages contain the code and output for answering each of the questions above for Example 6.6 on pages 264-267. The write up assumes that the desired family-wise error rate is $\alpha_T=0.05$.

Enter the Data from Table 6.21

The seven treatment groups can be broken into four experimental diets that contain a basal compound and:

cafo_1 corn and fish oil in a 1:1 ratio
calo_2 corn and linseed oil in a 1:1 ratio
faso_3 fish and sunflower oil in a 1:1 ratio, or
falo_4 fish and linseed oil in a 1:1 ratio

And three standard diets that were:

bc_5   the basal compound diet
lma_6 live micro algae
lmaa_7 live micro algae and *Artemia*


```
DATA shrimp_weights;
INPUT diet $ weight @@;
CARDS;
cafo_1 47.0      cafo_1 50.9      cafo_1 45.2      cafo_1 48.9      cafo_1 48.2
calo_2 38.1      calo_2 39.6      calo_2 39.1      calo_2 33.1      calo_2 40.3
faso_3 57.4      faso_3 55.1      faso_3 54.2      faso_3 56.8      faso_3 52.5
falo_4 54.2      falo_4 57.7      falo_4 57.1      falo_4 47.9      falo_4 53.4
bc_5   38.5      bc_5   42.0      bc_5   38.7      bc_5   38.9      bc_5   44.6
lma_6  48.9      lma_6  47.0      lma_6  47.0      lma_6  44.4      lma_6  46.9
lmaa_7 87.8      lmaa_7 81.7      lmaa_7 73.3      lmaa_7 82.7      lmaa_7 74.8
;
```

Check the Assumptions Using PROC INSIGHT and the Modified Levene test

```
PROC INSIGHT;
OPEN shrimp_weights;
FIT weight=diet;
RUN;

PROC GLM DATA=shrimp_weights ORDER=DATA;
CLASS diet;
MODEL weight=diet;
MEANS diet / HOVTEST=BF;
RUN;
```
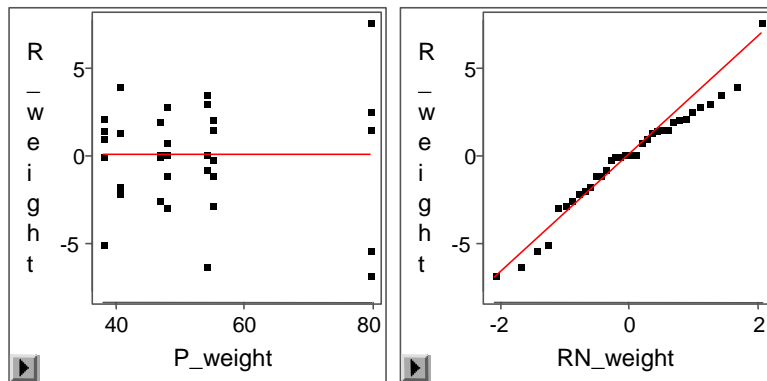

The GLM Procedure

Brown and Forsythe's Test for Homogeneity of weight Variance
ANOVA of Absolute Deviations from Group Medians

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
diet	6	41.4549	6.9091	1.32	0.2793
Error	28	146.1	5.2166		

From the residual vs. predicted plot, the means for each of the three groups seem to be near zero (in fact, they must always be for a one-way ANOVA). However, it is not clear from the residual vs. predicted plot if the variances of the errors for the seven groups are the same. Using the modified Levene test we fail to reject that they are different with a p-value of 0.2793. Finally, from the Q-Q plot of the residuals it appears that the distribution of the errors is approximately normally distributed, although its arguable at the ends.

Since it is a randomized experiment the assumption of independence is also satisfied so that all four assumptions for the ANOVA are met in this case.

Possible Question 1: Are there any differences between any of the group means?

```
PROC GLM DATA=shrimp_weights ORDER=DATA;
CLASS diet;
MODEL weight=diet;
RUN;
```

The GLM Procedure

Dependent Variable: weight

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	5850.774857	975.129143	88.14	<.0001
Error	28	309.792000	11.064000		
Corrected Total	34	6160.566857			

As the p-value of <.0001 is less than 0.05 we reject the null hypothesis

$$H_0{:}\mu_{cafo1}= \mu_{calo2} = \mu_{faso3}= \mu_{falo4} = \mu_{bc5}= \mu_{lma6}= \mu_{lmaa7}$$

and conclude that at least one of the diets has an effect that differs from the other six.

Possible Question 2: One possible set of specific contrasts.

Say that we are interested in the seven specific comparisons discussed on page 265:

1. How the experimental diets compare to the standard diets on average. (`newold`)
2. How diets containing corn compare to those without corn. (`corn`)
3. How diets containing fish oil compare to those without. (`fish`)
4. How diets containing linseed oil compare to those without. (`lin`)
5. How diets containing sunflower oil compare to those without. (`sun`)
6. How diets containing micro algae compare to those without. (`mic`)
7. How diets containing *Artemia* compare to those without. (`art`)

The seven corresponding null hypotheses could be written as:

1. H_0: $(\mu_{cafo1}+\mu_{calo2}+\mu_{faso3}+\mu_{falo4})/4 = (\mu_{bc5}+\mu_{lma6}+\mu_{lmaa7})/3$ [(3   3   3   3 -4 -4 -4) / 12]
2. H_0: $(\mu_{cafo1}+\mu_{calo2})/2 = (\mu_{faso3}+\mu_{falo4}+\mu_{bc5}+\mu_{lma6}+\mu_{lmaa7})/5$ [(5   5 -2 -2 -2 -2 -2) / 10]
3. H_0: $(\mu_{cafo1}+\mu_{faso3}+\mu_{falo4})/3 = (\mu_{calo2}+\mu_{bc5}+\mu_{lma6}+\mu_{lmaa7})/4$ [(4 -3   4   4 -3 -3 -3) / 12]
4. H_0: $(\mu_{calo2}+\mu_{falo4})/2 = (\mu_{cafo1}+\mu_{faso3}+\mu_{bc5}+\mu_{lma6}+\mu_{lmaa7})/5$ [(-2   5 -2   5 -2 -2 -2) / 10]
5. H_0: $\mu_{faso3} = (\mu_{cafo1}+\mu_{calo2}+\mu_{falo4}+\mu_{bc5}+\mu_{lma6}+\mu_{lmaa7})/6$ [(-1 -1   6 -1 -1 -1 -1) /   6]
6. H_0: $(\mu_{lma6}+\mu_{lmaa7})/2 = (\mu_{cafo1}+\mu_{calo2}+\mu_{faso3}+\mu_{falo4}+\mu_{bc5})/5$ [(-2 -2 -2 -2 -2   5   5) / 10]
7. H_0: $\mu_{lmaa7} = (\mu_{cafo1}+\mu_{calo2}+\mu_{faso3}+\mu_{falo4}+\mu_{bc5}+\mu_{lma6})/6$ [(-1 -1 -1 -1 -1 -1   6) /   6]

It is important that the data was entered in the correct order, and that we use the `ORDER=DATA` command.

```
PROC GLM DATA=shrimp_weights ORDER=DATA;
CLASS diet;
MODEL weight=diet;
ESTIMATE 'newold'  diet   3   3   3   3 -4 -4 -4 / divisor=12;
ESTIMATE 'corn'    diet   5   5 -2 -2 -2 -2 -2 / divisor=10;
ESTIMATE 'fish'    diet   4 -3   4   4 -3 -3 -3 / divisor=12;
ESTIMATE 'lin'     diet -2   5 -2   5 -2 -2 -2 / divisor=10;
ESTIMATE 'sun'     diet -1 -1   6 -1 -1 -1 -1 / divisor=6;
ESTIMATE 'mic'     diet -2 -2 -2 -2 -2   5   5 / divisor=10;
ESTIMATE 'art'     diet -1 -1 -1 -1 -1 -1   6 / divisor=6;
RUN;
```

| | | Standard | | |
Parameter	Estimate	Error	t Value	Pr > |t|
newold	-6.9783333	1.13613379	-6.14	<.0001
corn	-12.3000000	1.24457222	-9.88	<.0001
fish	1.0633333	1.13613379	0.94	0.3573
lin	-8.0860000	1.24457222	-6.50	<.0001
sun	3.9366667	1.60673582	2.45	0.0208
mic	16.2740000	1.24457222	13.08	<.0001
art	32.9400000	1.60673582	20.50	<.0001

Since we have seven tests that we are conducting, we need to determine which p-values to compare them to. We could do this by hand as follows (note we have put the tests in order of p-value).

```
Test         T       P-value     Compare To              Conclusion
art        20.50     <.0001      0.05/7=0.0071           reject H₀
mic        13.08     <.0001      0.05/6=0.0083           reject H₀
corn       -9.88     <.0001      0.05/5=0.0100           reject H₀
lin        -6.50     <.0001      0.05/4=0.0125           reject H₀
newold     -6.14     <.0001      0.05/3=0.0167           reject H₀
sun         2.45     0.0208      0.05/2=0.0250           reject H₀
fish        0.94     0.3573            0.05              fail to reject H₀
```

We could simultaneously make the following conclusions then:
1. We have significant evidence to conclude that the average effect of the four experimental diets is different from the average effect of the three standard diets (we estimate it to be 6.98 lower)
2. We have significant evidence to conclude that the average effect of the two diets with corn oil is different from the average effect of the five that don't (we estimate it to be 12.3 lower)
3. We do not have significant evidence to reject that the average effect of the three diets with fish oil are the same as the average effect of the four diets that don't
4. We have significant evidence to conclude that the average effect of the two diets with linseed oil is different from the average effect of the five that don't (we estimate it to be 8.09 lower)
5. We have significant evidence to conclude that the average effect of the diet with sunflower oil is different from the average effect of the six that don't (we estimate it to be 3.94 higher)
6. We have significant evidence to conclude that the average effect of the two diets with micro algae is different from the average effect of the five that don't (we estimate it to be 16.27 higher)
7. We have significant evidence to conclude that the effect of the diet with *Artemia* is different from the average effect of the six that don't (we estimate it to be 32.94 higher)

PROC MULTTEST could also be used to produce the above t-test p-values <u>and</u> automatically adjust them with the Holm procedure (what SAS calls the "Stepdown Bonferroni" procedure). This procedure will not produce the means that go with the contrasts however. A sample of the code is contained below in answering the third possible question.

Possible Question 3: What is the order of the group means, and which are significantly different from each other?

We will use `PROC MULTTEST` to conduct this procedure. (In reading the output, note that SAS also calls the Holm test the Stepdown Bonferroni test). To compare the means for all of the groups, we first need to enter all of the contrasts for comparing the means. There will always be (k choose 2) of them. Because there are $k=21$ groups in this example, there will only be such contrasts. We could also have entered these in `PROC GLM`, but `PROC MULTTEST` will automatically adjust all of the α levels, so that we don't have to!

```
PROC MULTTEST DATA=shrimp_weights ORDER=DATA HOLM;
CLASS diet;
CONTRAST '1 vs. 2'   1 -1   0   0   0   0   0;
CONTRAST '1 vs. 3'   1   0 -1   0   0   0   0;
CONTRAST '1 vs. 4'   1   0   0 -1   0   0   0;
CONTRAST '1 vs. 5'   1   0   0   0 -1   0   0;
CONTRAST '1 vs. 6'   1   0   0   0   0 -1   0;
CONTRAST '1 vs. 7'   1   0   0   0   0   0 -1;
CONTRAST '2 vs. 3'   0   1 -1   0   0   0   0;
CONTRAST '2 vs. 4'   0   1   0 -1   0   0   0;
CONTRAST '2 vs. 5'   0   1   0   0 -1   0   0;
CONTRAST '2 vs. 6'   0   1   0   0   0 -1   0;
CONTRAST '2 vs. 7'   0   1   0   0   0   0 -1;
CONTRAST '3 vs. 4'   0   0   1 -1   0   0   0;
CONTRAST '3 vs. 5'   0   0   1   0 -1   0   0;
CONTRAST '3 vs. 6'   0   0   1   0   0 -1   0;
CONTRAST '3 vs. 7'   0   0   1   0   0   0 -1;
CONTRAST '4 vs. 5'   0   0   0   1 -1   0   0;
CONTRAST '4 vs. 6'   0   0   0   1   0 -1   0;
CONTRAST '4 vs. 7'   0   0   0   1   0   0 -1;
CONTRAST '5 vs. 6'   0   0   0   0   1 -1   0;
CONTRAST '5 vs. 7'   0   0   0   0   1   0 -1;
CONTRAST '6 vs. 7'   0   0   0   0   0   1 -1;
TEST mean(weight);
RUN;
```



```
                        The Multtest Procedure

                    Continuous Variable Tabulations

                                                      Standard
            Variable     diet      NumObs     Mean    Deviation

            weight       cafo_1       5      48.0400    2.1267
            weight       calo_2       5      38.0400    2.8754
            weight       faso_3       5      55.2000    1.9812
            weight       falo_4       5      54.0600    3.9017
            weight       bc_5         5      40.5400    2.6857
            weight       lma_6        5      46.8400    1.6009
            weight       lmaa_7       5      80.0600    5.9777
```

```
                                    p-Values
                                              Stepdown
                Variable    Contrast        Raw    Bonferroni

                weight     1  vs.  2     <.0001      0.0006
                weight     1  vs.  3     0.0020      0.0132
                weight     1  vs.  4     0.0079      0.0315
                weight     1  vs.  5     0.0013      0.0106
                weight     1  vs.  6     0.5729      1.0000
                weight     1  vs.  7     <.0001      <.0001
                weight     2  vs.  3     <.0001      <.0001
                weight     2  vs.  4     <.0001      <.0001
                weight     2  vs.  5     0.2447      0.7340
                weight     2  vs.  6     0.0003      0.0026
                weight     2  vs.  7     <.0001      <.0001
                weight     3  vs.  4     0.5922      1.0000
                weight     3  vs.  5     <.0001      <.0001
                weight     3  vs.  6     0.0005      0.0041
                weight     3  vs.  7     <.0001      <.0001
                weight     4  vs.  5     <.0001      <.0001
                weight     4  vs.  6     0.0019      0.0132
                weight     4  vs.  7     <.0001      <.0001
                weight     5  vs.  6     0.0057      0.0285
                weight     5  vs.  7     <.0001      <.0001
                weight     6  vs.  7     <.0001      <.0001
```

We could use the column labeled Raw, and compare the values to $\alpha_T/21$, $\alpha_T/20$, etc... after putting them in order. Instead, however, the column labeled Stepdown Bonferroni has already been adjusted so that we can just compare those values directly to α_T. Thus, we would reject the null hypotheses that the first and second group have the same mean, but accept (fail to reject) the null hypothesis that the first and sixth group have the same mean. Notice that at α=0.05 there are only three hypotheses we fail to reject, but there are many more at α=0.01.

What way of presenting this would be to list the groups in order of their means, and have them share letters if they cannot be said to be significantly different.

```
Group        Mean Cortisol  Groupings at a=0.05    Groupings at a=0.01
lmaa_7       80.0600              A                       A

faso_3       55.2000                    B                       B
                                        B                       B
falo_4       54.0600                    B                  C    B
                                                          C    B
cafo_1       48.0400              C                       C    B    D
                                  C                       C         D
lma_6        46.8400              C                       C         D
                                                                    D
bc_5         40.5400                 D                    E         D
                                     D                    E
calo_2       38.0400                 D                    E
```

Note: Notice it is possible in some cases to have the groups overlap!! Remember, if we fail to reject we are not saying the means are equal, merely that we don't have enough evidence to say which are different for sure (α=0.01 means you need more evidence to say they are different than α=0.05).