Contrasts and Multiple Comparisons

Supplement for Pages 302-323

Brian Habing – University of South Carolina
Last Updated: July 20, 2001

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 α=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 α=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]

= 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]

= 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]

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

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

$= 1 - 0.95^{20}$

≈ 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:

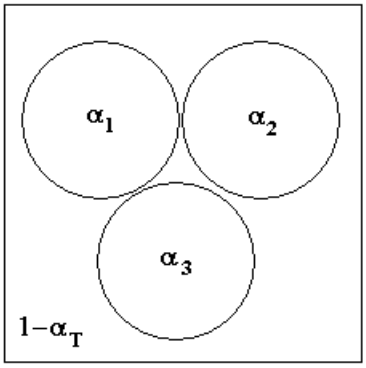
$$\alpha = 1 - (1-\alpha_T)^{1/k}$$

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!

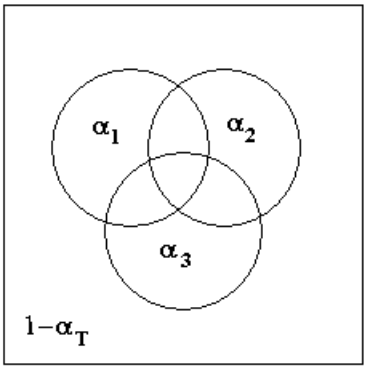
Bonferroni's Formula

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!

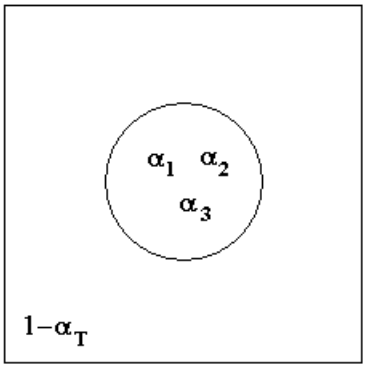
The diagram below illustrates three possible situations that could occur for three tests:



α_T is as large as possible
assume the worst
Bonferroni



α_T is in between
what usually happens
????



α_T is as small as possible
assume the best
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. 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 $\alpha=\alpha_T/(k\text{-}1)$. Again, if you fail to reject the null hypothesis then you stop here; if you do reject continue on and use $\alpha=\alpha_T/(k\text{-}2)$. (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 H_0 for test 5
1	0.011	0.05/4=0.0125	reject H_0 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 307-313), 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*.

A contrast is simply a linear function of the various treatment/group means whose coefficients sum to zero. Consider the example presented in Table 7-4 on pages 296-298. Here we have three groups: 1=healthy, 2=nonmelancholic-depressed, and 3=melancholic-depressed. Each of these three groups has an associated parameter: $\mu_1=\mu_h$, $\mu_2=\mu_{nonm\text{-}dep}$, and $\mu_3=\mu_{m\text{-}dep}$. Examples of contrasts here would include:

$L_1 = 0\cdot\mu_1 + 1\cdot\mu_2 - 1\cdot\mu_3$	written in SAS as	0 1 -1
$L_2 = 1\cdot\mu_1 + 0\cdot\mu_2 - 1\cdot\mu_3$		1 0 -1
$L_3 = 1\cdot\mu_1 - 1\cdot\mu_2 + 0\cdot\mu_3$		1 -1 0
$L_4 = 1\cdot\mu_1 - \frac{1}{2}\cdot\mu_2 - \frac{1}{2}\cdot\mu_3$		1 -0.5 -0.5

Notice that in each case the coefficients sum to 0: 0+1-1=0, 1+0-1=0, 1+0-1=0, 1-½-½=0. The theory says that we can estimate the contrasts using:

$$\hat{L} = \sum_{i=1}^{k} a_i \bar{y}_i \qquad \text{and} \qquad \hat{\sigma}_{\hat{L}} = \sqrt{MS_{res} \sum_{i=1}^{k} \frac{a_i}{n_i}}$$

where the a_i are the coefficients for the contrast, and the estimate $\hat{L}$ 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. The question though, is "why would we want to?"

If we look at L_1 we simply have the difference of the means of the second and third groups (the non-melancholic depressed and the melancholic depressed). It thus appears as if the contrast L_1 is simply comparing the means of those two groups. If we use the estimate of L and its standard error to construct a t-test of the hypothesis L_1=0, we get:

$$t_{df=N-k} = \frac{\bar{y}_2 - \bar{y}_3}{\sqrt{\frac{MS_{res}}{n_2} + \frac{MS_{res}}{n_3}}}$$

This is exactly the two-sample t-test for H_0: $\mu_1-\mu_2=0$ except that we are using MS_{res} instead of the pooled variance estimate! There is also an F-test for this contrast that tests exactly the same hypothesis (the F-value will always be the square of the t-value.)

If we return to the other three contrasts, L_2 is simply testing whether the non-depressed and melancholic depressed differ. Similarly L_3 is simply testing whether the healthy and nonmelancholic-depressed differ. The last contrast is somewhat more complicated. It is comparing the mean of the healthy to the average of the means of the two depressed groups. That is, it is comparing non-depressed to depressed.

Independence, Orthogonal Contrasts and the Holm-Sidak Test (an aside-note)

Two contrasts are said to be orthogonal if the dot-product of their coefficient vectors is zero. So, two contrasts $L = \sum_{i=1}^{k} a_i \mu_i$ and $L = \sum_{i=1}^{k} b_i \mu_i$ would be orthogonal if $\sum_{i=1}^{k} a_i b_i = 0$. In the above example then, L_1 and L_4 would be orthogonal, but no other pair of these contrasts would be. The reason to care if two contrasts are orthogonal is that the estimates that go with a set of orthogonal contrasts are independent. The test statistics will not be, however, as they both contain the MS_{res} in the denominator. There is a modification of the Holm test that uses Sidak's formula instead of Bonferroni's. However, because the statistics won't be independent, and there really isn't much difference between the values given by Bonferroni's formula and Sidak's formula, we will just use the basic Holm test.

Tying it All Together

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

1. A**re 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, something called Scheffé's method 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 - Hormones and Depression

The follow pages contain the code and output for answering each of the questions above for the example on pages 296-298. The write up assumes that the desired family-wise error rate is α_T=0.05.

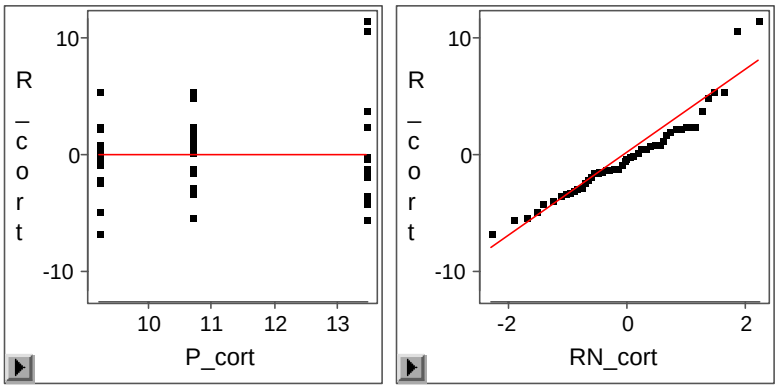
Enter the Data from Table 7.4

```
DATA tab7p4;
INPUT group $ cort @@;
CARDS;
h     2.5   n     5.4   m     8.1
h     7.2   n     7.8   m     9.5
h     8.9   n     8.0   m     9.8
h     9.3   n     9.3   m     12.2
h     9.9   n     9.7   m     12.3
h     10.3  n     11.1  m     12.5
h     11.6  n     11.6  m     13.3
h     14.8  n     12.0  m     17.5
h     4.5   n     12.8  m     24.3
h     7.0   n     13.1  m     10.1
h     8.5   n     15.8  m     11.8
h     9.3   n     7.5   m     9.8
h     9.8   n     7.9   m     12.1
h     10.3  n     7.6   m     12.5
h     11.6  n     9.4   m     12.5
h     11.7  n     9.6   m     13.4
            n     11.3  m     16.1
            n     11.6  m     25.2
            n     11.8
            n     12.6
            n     13.2
            n     16.3
;
```

Check the Assumptions Using PROC INSIGHT and the Modified Levene's test

```
PROC INSIGHT;
OPEN tab7p4;
FIT cort=group;
RUN;

PROC GLM DATA=tab7p4 ORDER=DATA;
CLASS group;
MODEL cort=group;
MEANS group / HOVTEST=BF;
RUN;
```



The GLM Procedure

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

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
group	2	6.5816	3.2908	0.48	0.6234
Error	53	365.9	6.9029		

From the residual vs. predicted plot, the means for each of the three groups seem to be near zero (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 three groups are the same. Using the modified Levene test we fail to reject that they are different with a p-value of 0.6234. Finally, from the Q-Q plot of the residuals it appears that the distribution of the errors is approximately normally distributed with the possible exception of two outliers.

Assuming the experimental design satisfies the independence assumption, then 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=tab7p4 ORDER=DATA;
CLASS group;
MODEL cort=group;
RUN;
```

The GLM Procedure

Dependent Variable: cort

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	164.6742857	82.3371429	6.61	0.0027
Error	53	660.0200000	12.4532075		
Corrected Total	55	824.6942857			

As the p-value of 0.0027 is less than 0.05 we reject the null hypothesis that $\mu_h = \mu_{nonm\text{-}dep3} = \mu_{m\text{-}dep}$ and conclude that at least one of the group cortisol means is different from the other two.

Possible Question 2: One example of specific contrasts.

Say that we have two research questions:

1. We want to know if the mean cortisol level of the healthy patients differs from the average of the means of the two types of depressed individuals.

2. We want to know if the mean cortisol level of the nonmelancholic depressed individuals differs from the mean level for the melancholic individuals.

The two null hypotheses would thus be

Hypothesis 1- H_0: $\mu_h = (\mu_{nonm\text{-}dep3} + \mu_{m\text{-}dep})/2$
Hypothesis 2- H_0: $\mu_{nonm\text{-}dep3} = \mu_{m\text{-}dep}$

These would correspond to contrasts 4 and 1 above. Note that it is important that the data was entered in the correct order, and that we use the `ORDER=DATA` command.

```
PROC GLM DATA=tab7p4 ORDER=DATA;
CLASS group;
MODEL cort=group;
ESTIMATE 'h vs. (n and m)'  group 1 -0.5 -0.5;
ESTIMATE 'n vs. m'          group 0  1   -1;
RUN;
```

Parameter	Estimate	Standard Error	t Value	Pr > |t|
h vs. (n and m)	-2.90000000	1.04537094	-2.77	0.0076
n vs. m	-2.80000000	1.12156130	-2.50	0.0157

As we have two tests, we compare the smallest p-value to $\alpha_T/2=0.05/2=0.025$. In this case it is the p-value for healthy vs. depressed, and the p-value of 0.0076 is smaller. We thus reject the null hypothesis for this test and conclude that the mean cortisol level for healthy individuals is different from the average of the mean cortisol levels for the two types of depressed individuals. In fact, we can estimate this difference and find that the healthy tend to have a cortisol level of 2.9 less than the average of the means for the other two groups. (Compare to the bottom of Table 7-4 on page 297).

Since we rejected the null hypothesis for the smallest p-value, we continue to the second of the two tests. We now compare the p-value of 0.0157 to $\alpha_T/1=0.05/1 = 0.05$. We can thus conclude that the nonmelancholic and melancholic individuals have significantly different mean cortisol levels. We estimate that the nonmelancholic individuals level is 2.8 lower than that of the melancholic individuals.

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 only *k*=3 groups in this example, there will only be 3 such contrasts. For k=10, there would have been 45. 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=tab7p4 ORDER=DATA HOLM;
CLASS group;
CONTRAST 'h vs. n'  1 -1 0;
CONTRAST 'h vs. m'  1 0 -1;
CONTRAST 'n vs. m'  0 1 -1;
TEST mean(cort);
RUN;
```

Continuous Variable Tabulations

Variable	group	NumObs	Mean	Standard Deviation
cort	h	16	9.2000	2.9305
cort	n	22	10.7000	2.7584
cort	m	18	13.5000	4.6742

p-Values

Variable	Contrast	Raw	Stepdown Bonferroni
cort	h vs. n	0.2014	0.2014
cort	h vs. m	0.0008	0.0025
cort	n vs. m	0.0157	0.0314

We could use the column labeled Raw, and compare the values to $\alpha_T/3$, $\alpha_T/2$, and $\alpha_T/1$. 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 for healthy vs. melancholic, and nonmelancholic vs. melancholic. We do not, however, have significant evidence to reject the null hypothesis for healthy vs. nonmelancholic.

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
Healthy	9.2	A
Nonmelancholic Depressed	10.7	A
Melancholic Depressed	13.5	B

Note: It is possible in some cases to have the groups overlap!! If nonmelancholic were more in the middle, we might not be able to say for certain it differed from either healthy or melancholic, even though melancholic and healthy clearly differed.