

Comparisons Between Treatment Means in ANOVA

*Dr. Tom Pierce
Department of Psychology
Radford University*

Let's say that a researcher wants to test the effects of an independent variable (let's say, the amount of tutoring) on a dependent variable (achievement test scores) and that the independent variable has three levels. The researcher decides to use Analysis of Variance to do this and does an F-test. As discussed in the last chapter this F-test is designed to answer the question "Are there any differences among the three sample means". This is a yes or no question. Either all three sample means are estimates of the same population mean or they're not. The researcher finds that this F-test is significant so they conclude that there is a significant effect of tutoring on achievement test scores.

Okay, so what do you know so far? You know that the independent variable has an effect on the dependent variable. You know that if you change the amount of tutoring you change the scores on the achievement test. That's something. But this is a very general piece of information. The F-test doesn't tell you whether people who get a lot of tutoring perform better than people who only get some tutoring. It doesn't tell you whether people who get some tutoring perform better than people who get no tutoring. It only tells you whether or not there is **any** effect of tutoring on achievement test scores. It gives you no information about **what** tutoring does to the scores. Because this F-test allows the researcher to make a decision about whether there are differences among any of the three sample means it's usually referred to as an **overall F-test** or an **omnibus F-test**.

When the overall F-test is significant, you've only completed the first step in figuring out what's going on with your data. Now, in order to say exactly what tutoring does to people's scores, you've got to determine which groups are different from which other groups. You've got to be able to perform a set of **comparisons** among the treatment means in the study.

This overall F-test is only the first step. In order to get more specific information about which groups are different from which other groups, the researcher needs to perform an additional set of tests to see which of these differences are significant. These tests are referred to as **comparisons among treatment means**. For example, one interesting comparison might test the difference between the mean for people who get no tutoring and the mean for people who get some tutoring. If you've got three treatment means there are three possible comparisons of one mean with another mean: Group 1 vs Group 2, Group 2 versus Group 3, and Group 1 versus Group 3.

Planned versus unplanned comparisons

One important distinction when conducting a set of comparisons is whether the comparisons are considered to be planned or unplanned. This is really a question of

whether or not the investigator has intended to test a certain comparison before the data were collected. Comparisons that the researcher intended to make before they collected the data are referred to as *planned comparisons*. Comparisons that the researcher decides to make after they get the data are referred to as *unplanned comparisons*.

The reason the distinction is important is because the techniques for testing planned and unplanned comparisons are different. Although we'll get into greater detail on this later, the basic issue here concerns the consequences of conducting a large number of statistical tests. I mean, think about it. When you test the difference between two means, what is the risk of making a Type I error? If the researcher uses an alpha level of .05 they're saying that they're willing to accept a five percent chance of making a Type I error. Okay, so how about if the researcher does three comparisons. What's the risk of making a Type I error somewhere in that set of comparisons? Well, there's a five percent chance of making a Type I error every time the researcher does a test, so the risk of making a Type I error anywhere in the set would have to be the number of comparisons (three) multiplied by the risk of making a Type I error for each comparison (five percent). This tells us that the risk of making a Type I error somewhere in that set of three comparisons is not the comfortably modest value of 5%, it's really 15%!

It turns out that there's a difference between the risk of making a Type I error when conducting a single comparison and the risk of making a Type I error anywhere in a set of comparisons. The risk of making a Type I error when testing a single comparison is referred to as the *per comparison alpha level*. The risk of making a Type I error anywhere in a set of comparisons is referred to as the *family-wise alpha level*. The different methods for conducting planned and unplanned comparisons allow the researcher some choice in terms of how they want to handle the problem of taking on added risk of a Type I error with every additional comparison the researcher performs. We'll talk about methods for conducting planned comparisons first.

Methods for conducting planned comparisons

The investigator in the tutoring and achievement test scores example has found that there is a significant overall effect of tutoring on achievement test scores. Anticipating this significant overall effect, let's say that the investigator made two predictions before they collected their data. First, they predicted that students who get a lot of tutoring will have significantly higher scores than students who get a moderate amount of tutoring. Second the investigator predicted that students who get tutoring, regardless of the amount, will have significantly higher scores on the achievement test than students who did not get tutoring.

Independent samples t-test

Without question, a researcher should be allowed to conduct a number of planned comparisons without have to make any type of adjustment for family-wise error. In other words, the researcher deserves an answer to the question of why the overall effect was significant and it's going to take several comparisons to address that question. A

family-wise alpha level of .10 or .15 is just the price that has to be paid for this information. If the researcher has two to four planned comparisons in mind, they should just go ahead and test these comparisons as regular old t-tests. There's no need to make it more difficult to reject the null hypothesis for any of these t-tests. In SPSS the way to get the results for these t-tests is through the Contrasts option that is placed at the bottom of the One-Way ANOVA window. You will often see comparisons reported as t-tests in results sections

Contrasts reported as F-tests

An alternative to reporting comparisons between treatment means as t-tests is to report them as F-tests. If you think for a second, what's the difference between comparing the means of groups 1 and 2 using a t-test and testing the difference between the means for these two groups using an F-test? No Difference! When there are only two groups, an F-test gives you exactly the same information that a t-test does. You can see this clearly from the probability levels for the F- and t-tests. They're identical, which means that they're both equally likely to yield a significant effect. Doing a t-test is the same thing as doing an F-test. One is no better or worse than the other one, although the F-test route carries a bit more flexibility in terms of the types of questions you can ask, as we'll see shortly.

The relationship between a value for t and a value for F is a very simple one. Let's say that an investigator has an independent variable and they test the difference between the two means using a t-test and then an F-test. As we mentioned in the previous paragraph the probability levels will be the same and the value for F will be equal to the value for t that has been squared ($F = t^2$). You'll often see comparisons between treatment means reported as F-tests. If you were doing your comparisons using F-tests in real life, you'd probably use a program like SPSS to do the calculations for you.

Let's say that you wanted to do the F-test for a comparison by hand. One reason for showing you this is that it helps to show you what the job of a set of comparisons really is.

We've already talked our way through the ANOVA table for the overall effect. There's a Sum of Squares Between-Groups (accounted for) and a Sum of Squares Within-Groups (not accounted for). The Within-Groups Sum of Squares represents variability that we don't have an explanation for. The Sum of Squares Between-Groups represents variability among the scores that we *attribute* to the effect of changing the level of the independent variable as we go from one group to the next. It represents the variability among all of the groups. This sum of squares accounted-for is something that we can take apart. There's a certain amount of variability that we can attribute to a difference between the scores in group 1 and the scores in group 2. This later amount of variability is a *part* of the total amount of variability accounted for. And we can test this specific amount of variability to see if it's significantly greater than what we might expect to get just by chance. In other words, is the variance (mean square) accounted for by a difference between the means for groups 1 and 2 significantly greater than the variance

(mean square) not accounted for (mean square within-groups)? All we have to do is to calculate a sum of squares for the comparison we're doing, then divide it by the appropriate number of degrees of freedom to get the mean square for that comparison. Then we take the mean square for the comparison and divide it by the mean square within-groups (that we already have) to get an F-ratio. One way of help think about this process is to just add an extra row to the ANOVA table we generated in the last chapter. The only difference is that this extra row will be dedicated to calculating an F-ratio for our comparison.

Source	SS	df	MS	F _{Observed}	F _{Critical}
Between	250	2	125	50	3.89
a1 vs a2	?	?	?	?	?
Within	30	12	2.5		

All right, so now let's calculate the sum of squares for the comparison. This process starts by calculating a value for D, according to the notation in Howell (2002). One way of thinking about this value for D is that it basically represents the difference between the means being compared. A value of zero for D indicates that there isn't any difference at all between the means. The further away the value for D is from zero, the bigger the difference between the means. To calculate the value for D you first list the means for all of the levels of the independent variable in order. So the means for groups one, two, and three are ...

13 8 3

The next step is to multiply each mean by a weighting or **coefficient** that reflects the contribution of that mean to the comparison being made.

13() 8() 3()

The rules for generating a set of coefficients for a particular comparison are that (a) the coefficients have to add up to zero and (b) the pattern of the coefficients as you move across the various levels of the IV have to reflect or **code** the comparison being made. One should be able to look at the coefficients multiplied by the means and know what the comparison is. In terms of trying to explain how to generate these coefficients, it's easier just to show you through a couple of examples than to define some elaborate set of rules that makes it harder than it really is. Here goes.

We want to compare the mean of Group 1 to the mean of Group 2. Does Group 3 have anything to do with this comparison? No. Okay, so how much does this group contribute

to the comparison? Nothing. Okay, so what coefficient do you think the mean for this group ought to get? Zero! Right. So the mean of 3 gets multiplied by a zero.

$$13() \quad 8() \quad 3(0)$$

Next, you know that the coefficients have to add up to zero. So if you make the coefficient applied to group 1 equal to +1, what are you going to have to make the coefficient for group 2? It's going to have to be -1. So now we've got...

$$13(+1) \quad 8(-1) \quad 3(0)$$

The set of coefficients -1, +1, 0 is said to "code" the comparison of the mean of group 1 to the mean of group 2. Now, to generate the value for D that we need, all you have to do is to multiply the means by their coefficients and then add these numbers up.

$$+13 - 8 + 0 = +5$$

The value for D is +5. Obviously, there is a five point difference between the mean for group 1 and the mean for group 2. Next, we take this number and plug it into an equation that gives us the sum of squares for the comparison.

$$SS_{\text{Comparison}} = \frac{n(D^2)}{\sum c^2}$$

The top part of the equation is easy. "n" refers to the number of people in each group, which is 5. So "n(D²)" becomes "5(5²)" = 5(25) = 125. The bottom part of the equation "Σc²" refers to the number you get when you take each of the coefficients, square them, and then add these squared numbers up. So here we've got "0² + (+1)² + (-1)² = 0 + 1 + 1 = +2. So the number crunching for the equation ends up looking like this...

$$SS_{\text{Comparison}} = \frac{n(D^2)}{\sum c^2} = \frac{5(5^2)}{+2} = \frac{125}{2} = 67.5$$

The sum of squares for this particular comparison is 67.5. Now let's plug it into the ANOVA table and see what happens.

Source	SS	df	MS	F _{Observed}	F _{Critical}
Between	250	2	125	50	3.89
a1 vs a2	67.5	?	?	?	?
Within	30	12	2.5		

We want the F-ratio for the comparison. We've got the sum of squares. How many degrees of freedom should we divide this sum of squares by? Well, how many groups are involved in the comparison? Two. And what's the equation for determining the number of degrees of freedom for the accounted-for term? It's the number of levels of the independent variable minus one ($a - 1$). If we've got two means being compared, then two levels minus one leaves us with one degree of freedom. The number of degrees of freedom for the comparison is one. It turns out that because *every* comparison is the comparison of one group of scores to another group of scores, every comparison has one degree of freedom associated with it.

If the comparison has one degree of freedom, that means that the Mean Square for the comparison is equal to 67.5 divided by one, which just leaves us with 67.5.

Source	SS	df	MS	F _{Observed}	F _{Critical}
Between	250	2	125	50	3.89
a1 vs a2	67.5	1	67.5	?	?
Within	30	12	2.5		

The last step is to determine the F-ratio for the comparison. The F-ratio for the comparison is computed by taking the mean square for the effect being tested (the comparison) and dividing it by the mean square for the error term (the sum of squares within-groups) which is 2.5. This gives us an F-ratio of 25

Source	SS	df	MS	F _{Observed}	F _{Critical}
Between	250	2	125	50	3.89
a1 vs a2	67.5	1	67.5	25	4.75
Within	30	12	2.5		

The only thing we need to know now is the critical value used to test this F-ratio for significance. Just like before, the critical value for F is based on the number of degrees of freedom for the numerator and the number of degrees of freedom in the denominator. Here we need to look up the critical value for F when there is one degree of freedom in the numerator and 12 degrees of freedom in the denominator. The number we get is 4.75.

The observed value of 25 is greater than the critical value of 4.75, so our decision is to reject the null hypothesis that there is no difference between the two means. Our conclusion is that "The mean of students getting a lot of tutoring is significantly greater than the mean of students getting a moderate amount of tutoring, $F(1, 12) = 25, p < .05$."

Now, on to the second planned comparison of people who get tutoring compared to people who do not get tutoring. Who's getting compared in this comparison? Remember, any comparison is between two groups of people. Obviously, there are two different groups who got tutoring (lot of tutoring, some tutoring) and one group that didn't get tutoring. In this comparison it doesn't matter how much tutoring people get. We're really comparing the average score for everybody who got tutoring (i.e., people in a1 or a2) to the average score for people who didn't get tutoring (a3). As shorthand, you could represent this comparison as (a1 + a2 vs. a3). This is an example of a **complex comparison** because more than one group is represented on at least one side of the comparison. The first comparison (a1 vs. a2) is referred to as a **pair-wise comparison** because only two – a pair – of means are involved in the comparison.

So how do you get the coefficients? Remember that every comparison has two sides to it. Also, the coefficients have to add up to zero. For whatever amount of weight you give to the positive side of the comparison, you'll have to give the same amount of weight to the negative side of the comparison. So, if you make the coefficients on the positive side add up to +2, you've got to have the negative coefficients add up to -2. We've got two means on the positive side of the comparison (a1 and a2) and we want to give equal weight to both means so we could assign a coefficient of +1 to a1 and a coefficient of +1 to a2. That way the total for the two coefficients comes up to +2. The only mean on the negative side of the comparison is a3. So we could assign a coefficient of -2 to a3. That makes the entire set of coefficients (+1, +1, -2). When we use these coefficients to calculate the value of D for this second comparison we get...

$$D = 13(+1) + 8(+1) - 3(-2) = 13 + 8 - 6 = 21 - 6 = 15$$

Now we take this value for D of 15 and plug it into the equation for the sum of squares for a comparison.

$$SS_{\text{Comparison}} = \frac{n(D^2)}{\sum c^2} = \frac{5(15)^2}{6} = \frac{5(225)}{6} = \frac{1125}{6} = 187.5$$

In the denominator of the equation we need to calculate the sum of the squared coefficients, which is: $1^2 + 1^2 + 2^2 = 1 + 1 + 4 = 6$. So the sum of squares for the second comparison is 187.5. When we put add this second comparison to the ANOVA table we get...

Source	SS	df	MS	F _{Observed}	F _{Critical}
Between	250	2	125	50	3.89
a1 vs a2	67.5	1	67.5	25	4.75
a1 + a2 vs. a3	187.5	1	187.5	75	4.75
Within	30.0	12	2.5		

This second comparison has 1 degree of freedom. This may seem strange because all three groups are involved in the comparison, but it's still the case that this comparison is testing the mean of one group of people (people who got tutoring) against the mean of a second group of people (people who did not get tutoring). Two levels of the independent variable minus one degree of freedom leaves you with one degree of freedom. The critical value for this second comparison stays the same because the degrees of freedom for the numerator and the denominator of the F-ratio stay the same. So this second comparison is significant and we can conclude that "Students who get tutoring, regardless of the amount, perform significantly better than students who do not get tutoring, $F(1,12) = 75.0$. $p < .05$."

Orthogonal versus non-orthogonal comparisons

Now take a look at the sums of squares for the two comparisons. When you add them up, what do you get? 250! That's the sum of squares accounted for! How did that happen? It turns out that **when the independent variable has three groups, it only takes two comparisons to provide an explanation for where the sum of squares accounted-for came from**. It's no accident that this number of comparisons corresponds to the number of degrees of freedom for the overall effect. The combined information from these two comparisons is able explain why the overall effect was significant. Together they explain exactly as much variability as there was to explain, no more and no less.

Now, let's say that we wanted to conduct a third comparison that compares the mean of people who get a lot of tutoring to the mean of people who get some tutoring or no tutoring. This is the comparison of a_1 versus a_2 and a_3 . Coefficients to test this comparison might be "+2, -1, -1". When you plug these coefficients into the Contrasts option in SPSS, get the observed value for t, square it to get the observed value for F, and then figure out the sum of squares for this comparison you get 187.5.

When you add the sums of squares for comparisons 1 and 2 together you get exactly 250. When you add the sums of squares for comparisons 2 and 3 together ($187.5 + 187.5$) you get a number that's a lot greater than 250. It looks like comparisons 2 and 3, taken together, account for more variability than there was to account for in the first place. Why is that? It's because of a distinction that one can make among sets of comparisons. This distinction is in terms of whether the comparison are said to be **orthogonal** to each other or **non-orthogonal** to each other.

Two comparisons are orthogonal to each other when they don't overlap at all in terms of the information they provide. In other words, when the comparisons you're talking about are orthogonal they're answering completely separate questions about the effects of the independent variable on the dependent variable. Let's think about comparisons 1 and 2 for a second. In comparison 1, the only thing happening is that we're looking at the difference between groups 1 and 2. Group three has nothing to do with this first comparison. In the second comparison, we're not looking at the difference between groups 1 and 2. We're averaging over groups 1 and 2. We're treating the people in groups 1 and 2 like they come from the same group. And we're comparing these people

(people who got at least some tutoring) to people who didn't get any tutoring. In comparison 1, group 3 wasn't involved at all. In comparison 2, group 3 is the only level of the independent variable that is on one side of the comparison. Comparisons 1 and 2 are said to be orthogonal to each other because they are addressing completely separate questions.

On the other hand, think about what comparisons 2 and 3 are doing. What do comparisons 2 and 3 both have in common? They both have group 1 on one side of the comparison and group 3 on the other side of the comparison. In a very real sense, comparisons 2 and 3 are answering similar questions. They're not identical questions, but the two comparisons both provide information about the difference between the means for groups 1 and 3. That means that when you test comparisons 2 and 3, you're getting some of the same information twice. That's why, when you add up the sums of squares for comparisons 2 and 3, you get a number greater than 250. It's because, to some extent, you're testing the same questions twice.

Some statisticians think that researchers should be restricted to testing orthogonal comparisons. They feel this way because testing the same variability twice means that the researcher has two chances to make a Type I error when they test this variability. There are two chances to make the one mistake. That's a situation where the researcher is at double jeopardy of saying that there's something there when there really isn't. Other statisticians feel that it's okay to conduct non-orthogonal comparisons as long as these comparisons provide answers to interesting and meaningful questions.

One quick way to tell if two comparisons are orthogonal or not is to start by stacking the coefficients for these comparisons on top of each other. For comparisons 1 and 2 we'd have...

$$\begin{array}{r} +1 \ -1 \ 0 \\ +1 \ +1 \ -2 \end{array}$$

Now multiply the coefficients going down each column.

$$\begin{array}{r} +1 \ -1 \ 0 \\ +1 \ +1 \ -2 \\ \hline +1 \ -1 \ 0 \end{array}$$

After you've got these numbers at the bottom of each column, you just add these numbers up. $(+1) + (-1) + 0$ gives you a value of zero. Using this little trick, if you get zero the two comparisons are orthogonal. If you get anything other than zero the comparisons are not orthogonal. This method tells us that comparisons 1 and 2 are orthogonal to each other.

Now let's do the same thing for comparisons 2 and 3. When we stack the coefficients on top of each other we get...

$$\begin{array}{r} +1 \ +1 \ -2 \\ +2 \ -1 \ -1 \end{array}$$

Now multiplying the coefficients going down each column we get...

$$\begin{array}{r} +1 \ +1 \ -2 \\ +2 \ -1 \ -1 \\ \hline +2 \ -1 \ +2 \end{array}$$

Adding these numbers up, $(+2) + (-1) + (+2)$, gives you $+3$. This number is something other than zero, so we know that these two comparisons are not orthogonal to each other. This method can tell the researcher whether their comparisons are orthogonal to each other before they even collect their data.

Corrections for family-wise risk of a Type I error: The Bonferroni adjustment

So how many planned comparisons should a researcher be allowed to do? Obviously, each comparison answers a different question regarding the effect of the independent variable on the dependent variable. On the face of it, why should there be *any* limit to the number of questions that a researcher should be allowed to ask? It turns out that there is quite a bit of disagreement among statisticians on this. What I'm going to do is to tell you about the basic issue that the statisticians are wrestling with and then I'll describe a couple of approaches to addressing the issue.

Let's say that a researcher tests one planned comparison and uses an alpha level of .05. What are the chances that the researcher is going to commit a Type I error if they decide to reject the null hypothesis? Five percent, obviously. Okay. Now let's say that the researcher conducts a second planned comparison and uses an alpha level of .05. What are the odds that they've committed a Type I error if they reject the null hypothesis for this second comparison? Again, five percent. All right, now, in doing those two planned comparisons, what are the odds of committing a Type I error in *either* comparison? ***What are the odds of committing a Type I error in the set, or the family, of comparisons?*** Well, they did two comparisons and each comparison carried a five percent risk of making a Type I error. So, overall, the odds were 10% that the researcher would commit a Type I error in the set of two comparisons. It turns out that the more comparisons the researcher does the higher the overall risk of making a Type I error gets. There is an important distinction between the risk of making a Type I error for a single comparison and the risk of making a Type I error in a family or a set of comparisons.

The **per comparison alpha level** refers to the risk of making a Type I error for a single comparison. The **family-wise alpha level** refers to the risk of making a Type I error anywhere in a family or a set of comparisons.

The per comparison alpha level is the alpha level the researcher has decided to use to test an individual comparison. The size of family-wise alpha level is a function of two things: the alpha level used to test each individual comparison and the number of comparisons being tested. I'm going to show you two equations for the family-wise alpha level. The first equation gives the actual risk of making a Type I error in a set of comparisons. This equation is awkward to use but gives you the correct answer. The second equation only gives you an *approximation* of the actual family-wise alpha level. However, this second equation is much easier to use. Because answers using this second equation are very close to the real thing, especially when the number of comparisons is less than five or so, this is the equation we're going to work with.

The equation to calculate the actual family-wise risk of making a Type I error is

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

If the researcher decides to use a per comparison alpha level of .05 and they want to conduct three planned comparisons they'll end up with the following family-wise alpha level:

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

$$\alpha_{FW} = 1 - (1 - .05)^3$$

$$\alpha_{FW} = 1 - (.95)^3$$

$$\alpha_{FW} = 1 - .857 = .143$$

According to this equation the odds of committing a Type I error anywhere in the set of three planned comparisons is 14.3%.

Okay, here's the second equation. It's based on the same reasoning we talked our way through above. It basically says that if the per comparison alpha level is .05 then the researcher takes on 5% worth of risk for every comparison they do. This equation says that the family-wise alpha level is equal to the per comparison alpha level the researcher is using multiplied by the number of comparisons the researcher has decided to do. In symbol form...

$$\alpha_{FW} = (\alpha_{PC})(c) \text{ or, for our example...}$$

$$\alpha_{FW} = (.05)(3) = .15$$

According to this second, and far easier equation, the family-wise risk of a Type I error is 15%, which is pretty close to the actual value of 14.3%. If anything, this second equation slightly overestimates the actual family-wise alpha level.

Okay, so that's how a researcher can calculate the family-wise risk of making a Type I error. The critical question here is whether there should be a limit to the size of the family-wise alpha level. In other words, just how much total risk of making a Type I error is too much? Unfortunately, this is one of those questions where statisticians disagree. I'll give you two ways of handling the situation and leave it to you to decide which way makes the most sense to you.

A number of statisticians feel that researchers should be able to test as many comparisons as it takes to answer their important questions (e.g., Keppel, Saufley, & Tokunaga, 1992). If this means that the researcher tests five comparisons and ends up with a 25% chance of committing a Type error, then so be it. The benefits of getting answers to five important questions outweigh the potential costs that result from making one or more Type I errors. This is a rather liberal approach to statistical decision making.

A person taking a slightly more conservative approach to statistical decision making might make the following argument...

The job of a set of comparisons is to explain why the overall ANOVA was significant. It takes as many orthogonal planned comparisons to do this as you have degrees of freedom for the overall effect. Therefore, the researcher should be able to perform at least this many comparisons without having to worry that the family-wise risk of making a Type I error is getting too large. However, if the researcher wants to do more planned comparisons than the number of degrees of freedom for the overall effect, the researcher is now at the point where they have to worry that the family-wise alpha level is getting too high.

For example, if the independent variable has three levels then the researcher should be allowed to conduct two planned comparisons without any kind of penalty. This means that the researcher is allowed to let the family-wise alpha level get up as high as 0.10, but they can't let it get any higher than this. The highest they can let the family-wise alpha level get is thus...

$$\text{Maximum } \alpha_{FW} = (df_A)(\text{original } \alpha_{PC}) = (2)(.05) = 0.10$$

There is nothing wrong at all with wanting to test more than two planned comparisons in this example, but how could the researcher do this and still not let the family-wise alpha level get above 0.10? Let's say they want to test four planned comparisons. Obviously, if the researcher wants to test more than two planned comparisons they're going to have to use a lower per-comparison alpha level. In this case if the researcher used a per-comparison alpha level of 0.025 they could test their four planned comparisons and still keep the family-wise alpha level from going over 0.10. The **Bonferroni Adjustment** provides an equation for calculating an *adjusted per-comparison alpha level* to use when the number of planned comparisons is larger than the number of degrees of freedom for the overall effect. This adjusted per-comparison alpha level is basically just the maximum family-wise alpha level that the researcher is allowed to have divided by the number of planned comparisons the researcher actually wants to do.

$$\text{Adjusted } \alpha_{PC} = \frac{(\text{df}_A)(\text{original } \alpha_{PC})}{c}$$

The value for “c” in this case represents the number of comparisons. Applying the Bonferroni Adjustment to this example, we get...

$$\text{Adjusted } \alpha_{PC} = \frac{(2)(0.05)}{4} = \frac{0.10}{4} = 0.025$$

The strategy employed in the Bonferroni Adjustment is used widely in data analysis in situations where the researcher is faced with conducted a large number of different tests.

Unplanned Comparisons

When conducting unplanned comparisons the investigator is prepared to look in a large number of places to detect differences between means. The researcher needs to guard against the risk of making a Type I error anywhere in the set of all possible locations where you could look. Because the number of all *possible* comparisons is larger than a limited number of planned comparisons methods for testing unplanned comparisons make it far more difficult to reject the null hypothesis for any one of these comparisons. We'll look at two strategies for conducting unplanned comparisons, the Scheffe test and the Tukey test.

Scheffe Test

The Scheffe test is the most conservative method for conducting unplanned comparisons. It allows researcher to test every possible comparison, pair-wise and complex. SPSS provides a table that reports the results for all pair-wise comparisons conducted using the Scheffe method. Unfortunately, the critical value is adjusted based on the assumption that the investigator is testing every possible comparison, including all of the complex comparisons. So SPSS doesn't give you access to all of the comparisons you're paying for. For this reason, it doesn't make much sense look at SPSS's Scheffe output without being willing to test at least some of the complex comparisons through the method outlined above using the Contrasts option. SPSS's Tukey output gives you the same pair-wise comparisons, but tests them using a lower critical value.

$F_S = (\text{degrees of freedom for effect})(\text{Critical value for overall effect})$
 Plugging the number in from our example, the critical value for F for the Scheffe test becomes...

$$F_S = (2)(3.89) = 7.78$$

... a value of 7.78. You'll notice that this critical value is quite a bit higher than the one used to test a planned comparison (4.77). And, with four or five groups the critical value would go much higher than that. This is what makes the Scheffe test the most conservative method for testing comparisons among treatment means. This is what makes it the method of choice for the researcher who is especially concerned about committing a Type I error. With the Scheffe test, you've got permission to test any and all comparisons without having to worry about an inflated family-wise risk of a Type I error. One way of thinking about the Scheffe test is that it's the all-you-can-eat buffet of testing comparisons. You're paying a high price for being able to go back for as many F-tests as you can stomach.

Tukey method

The Tukey method assumes that the researcher is going to test all possible pairwise comparisons. A pairwise comparison is one treatment mean compared to one other treatment mean (e.g., a_1 vs a_2). The tests of unplanned comparisons using the Tukey method are conducted using an adjusted critical value for F. The more possible pair-wise comparisons there are, the larger this critical value will be. Everything about obtaining the observed value for F is the same as for a planned comparison. The only thing that changes is the critical value for F. The equation for calculating this adjusted critical value for F is...

$$F_T = \frac{q^2}{2}$$

The adjusted critical value for F for the Tukey Test (F_T) is equal to squaring the value for a statistic known as the Studentized Range (q) and then dividing by 2. It doesn't get much easier than that. The value for q can be found in the Critical Values for the Studentized Range table. To know which row to look in to find the value for q you need to know the number of degrees of freedom for the Sum of Squares within-groups (the denominator of the overall effect). To know which column to look in you need to know the number of levels of the independent variable. In the table we're using, the number of levels of the independent variable is referred to as a value for K . At 12 degrees of freedom for the sum of squares within-groups and three levels of the independent variable we get a value for q of 3.77. Plugging this number into the equation we get...

$$F_T = \frac{q^2}{2} = \frac{3.77^2}{2} = \frac{14.21}{2} = 7.11$$

... a value of 7.11 to use as the as the adjusted critical value for F when conducting unplanned comparisons using the Tukey method.

Take a moment to compare the adjusted critical values for F for the Scheffe and Tukey methods. The critical value for the Scheffe test is 7.78 and the critical value for the

Tukey method is 7.11. Why is the critical value for the Scheffe test higher than the critical value for the Tukey test? It's because the Scheffe test allows you to test more possible comparisons (i.e., every possible comparison, as opposed to every possible pairwise comparison). The more possible places there are where something could be significant just by chance, the more difficult you have to make it to reject the null hypothesis for each one of those tests.