Richard Gonzalez
Psych 613
Version 3.1 (Sep 2022)

LECTURE NOTES #3: Contrasts and Post Hoc tests

Reading assignment

- Read MD chs 4, 5, & 6
- Read G chs 7 & 8

Goals for Lecture Notes #3

- Introduce contrasts
- Introduce post hoc tests
- Review assumptions and issues of power

1. Planned contrasts

   (a) Contrasts are the building blocks of most statistical tests: ANOVA, regression, MANOVA, discriminant analysis, factor analysis, etc. We will spend a lot of time on contrasts throughout the year.

   A contrast is a set of weights (a vector) that defines a specific comparison over scores or means. They are used, among other things, to test more focused hypotheses than the overall omnibus test of the ANOVA.

   For example, if there are four groups, and we want to make a comparison between the first two means, we could apply the following weights: 1, -1, 0, 0. That is, we could create a new statistic using the linear combination of the four means

$$\hat{\Psi} = (1)\overline{Y}_1 + (-1)\overline{Y}_2 + (0)\overline{Y}_3 + (0)\overline{Y}_4 \quad (3\text{-}1)$$
$$= \overline{Y}_1 - \overline{Y}_2 \quad (3\text{-}2)$$

   This contrast is the difference between the means of groups 1 and 2 ignoring groups 3 and 4 (those latter two groups receive weights of 0).

The contrast $\hat{I}$ is an estimate of the true population value I. In general, the contrast $\hat{I}$ is computed by

$$\sum a_i \overline{Y}_i \tag{3-3}$$

where the a_i are the weights. You choose the weights depending on what research question you want to test.

The two sample *t* test is equivalent to the 1, -1 contrast on the two means (when the design includes only two groups). Consider the null hypothesis:

H_O: $\mu_1 - \mu_2 = 0$
H_O: $(1)\mu_1 + (-1)\mu_2 = 0$

Thus, the null hypothesis of the two sample t-test is that the 1, -1 contrast on the two population means is equal to zero.

Back to the example with four groups. Because we're assuming all four groups have equal population variances we can pool together the variances from the four groups to find the error term to test the contrast $\hat{I}$. This yields a more powerful test compared to completely ignoring the other two groups because information from all four groups is pooled to estimate the error term. But, as usual, the assumption of equal population variances is crucial. You can see how the test becomes meaningless if the variances are grossly discrepant. There is increasing sentiment among applied statisticians that contrasts should be tested with the separate variance test to avoid the homogeneity of variance assumption. SPSS output gives both the classic test for the contrast as well as a Welch-type correction that does not assume equal variances (labeled as the separate variance test in SPSS output). There hasn't been an easy way to test Welch-style contrasts in R so I wrote a little function to do this, which I'll describe later in these lecture notes.

(b) How to test the alternative hypothesis that the population I $\neq 0$?

Recall from the first set of lecture notes that

$$t \sim \frac{\text{estimate of population parameter}}{\text{st. dev. of sampling distribution}} \tag{3-4}$$

In the case of contrasts, Equation 3-4 becomes

$$t \sim \frac{\hat{I}}{\text{st. error}(\hat{I})} \tag{3-5}$$

From Equation 3-1 we already know how to compute $\hat{I}$ (which is the numerator in Equa-

tion 3-5). The standard error of $\hat{\text{I}}$ is also straightforward.

$$\text{st. error}(\hat{\text{I}}) \quad = \quad \sqrt{\text{MSE} \sum_{i=1}^{T} \frac{\text{a}_\text{i}^2}{\text{n}_\text{i}}} \tag{3-6}$$

where MSE is the mean square error term from the overall ANOVA on all groups (recall that MSE is the same as MSW for now) and the a_i are the contrast weights. The critical t is a "table look-up" with N - T degrees of freedom (i.e, the same degrees of freedom associated with the MSE term)[1]. Typically, you want a two-tailed test of the contrast. Most statistical packages report the two-tailed *p* value; a one-tailed test is also possible for those who want to report them.

The hypothesis test for $\hat{I}$ can be summarized using the hypothesis testing template introduced in Lecture Notes 1. See Figure 3-1.

It is possible to construct a confidence interval around $\hat{\text{I}}$

$$\hat{\text{I}} \quad \pm \quad t_{\alpha/2}\text{se}(\hat{\text{I}}) \tag{3-8}$$

The t in this confidence interval has the same degrees of freedom as MSE (i.e., N - T).

Note that the error term of the contrast (the denominator of the t) uses information from all groups regardless of whether the contrast weight is 0; whereas, the value of the contrast $\hat{\text{I}}$ (the numerator of the t) ignores means with a weight of zero. This is an important distinction!

(c) Orthogonality.

Orthogonality is defined between pairs of contrasts. Take two contrasts $(\text{a}_1, \text{a}_2, \ldots, \text{a}_\text{t})$ and $(\text{b}_1, \text{b}_2, \ldots, \text{b}_\text{t})$. If the sample sizes are equal and

$$\sum \text{a}_\text{i}\text{b}_\text{i} \quad = \quad 0, \tag{3-9}$$

then we say contrast A and contrast B are orthogonal. Orthogonality means that the two contrasts are not correlated (i.e., the covariance between A and B is zero).

[1] One could equivalently perform an F test by squaring Equation 3-5

$$\text{F}_{1,\ \text{df}_\text{error}} \quad = \quad \frac{\hat{\text{I}}^2}{\text{MSE} \sum \frac{\text{a}_\text{i}^2}{\text{n}_\text{i}}} \tag{3-7}$$

The degrees of freedom for the error are N - T (i.e., the number of subjects minus the number of groups). In the context of the F test, contrasts **always** have one degree of freedom in the numerator.

Figure 3-1: Hypothesis Testing Framework for Contrasts

Null Hypothesis

- H_o: $I = 0$
- H_a: $I \neq 0$ (two-sided test)

where $I = \sum a_i \overline{Y}_i$.

Structural Model and Test Statistic

The structural model follows the usual ANOVA model.

The test statistic operates on the weighted sum I and specifies its sampling distribution. The test of the hypothesis will involve an estimate over the standard error of the estimate, therefore we make use of the definition of the *t* distribution

$$t \sim \frac{\text{estimate of population parameter}}{\text{estimated st. dev. of the sampling distribution}}$$

Using the statistical results stated in the lecture notes we write the specific details for this problem into the definition of the *t* distribution

$$t_{\text{observed}} = \frac{\hat{I}}{\sqrt{\text{MSE} \sum_{i=1}^{T} \frac{a_i^2}{n_i}}}$$

with df = $N - T$, which is total sample size minus the number of groups in the design.

Critical Test Value We use the *t* table to find the critical value of *t*, denoted t_{critical} for the specific degrees of freedom, two-sided, and $\alpha = 0.05$.

Statistical decision If $|t_{\text{observed}}| > t_{\text{critical}}$, then reject the null hypothesis, otherwise fail to reject.

This definition applies only when there are equal sample sizes. When the sample sizes are unequal, orthogonality can be defined as

$$\sum \frac{a_i b_i}{n_i} = 0. \quad (3\text{-}10)$$

Orthogonality is defined over pairs of contrasts

A set of contrasts is said to be orthogonal if all possible pairs of contrasts within the set are orthogonal. So, if there are 4 contrasts labelled A, B, C and D, then all six possible pairs (AB, AC, AD, BC, BD, and CD) must each be orthogonal for the set of 4 contrasts to be orthogonal to each other[2].

We will discuss different ways that sample size enters how we define effects when we cover main effects in factorial ANOVAs. I tend to favor what is called the regression approach where orthogonality is defined only by the contrast weights (as in Equation 3-9) without consideration of sample size.

i. Making orthogonality intuitive

Orthogonality refers to the angle between the two contrasts. To understand this we must give a geometric interpretation to the contrast. A contrast is a point in a T-dimensional space, where T is the number of groups. For example, the contrast (1, 1, -2) is a point in a three dimensional space with values of 1 for the first coordinate (or axis), 1 for the second coordinate, and -2 for the third coordinate. Each contrast defines a vector by drawing a straight line from the origin of the space to the point. The angle between pairs of vectors becomes critical. If the two contrasts form a 90° angle, then they are orthogonal.

ii. An easy way to see orthogonality is to consider the simple pair of contrasts (1,1) and (1,-1). Check that these are orthogonal. We can plot these two contrasts and see that they are at 90 degrees from each other—see Figure 3-2.

There are many different sets of contrasts that satisfy orthogonality. For example, this set of three row contrasts for the case of four groups is orthogonal

$$\begin{array}{rrrr} 1 & -1 & 0 & 0 \\ 1 & 1 & -2 & 0 \\ 1 & 1 & 1 & -3 \end{array}$$

[2]Every year we see students doing strange things to examine orthogonality in a set of contrasts, such as multiply all individual contrasts together to determine orthogonality. Again, orthogonality in a set of contrasts is checked separately for every possible pair of contrasts. An unrelated concept, interactions, involves multiplying two or more contrasts together to create a new contrast. We will cover interactions in later lecture notes.

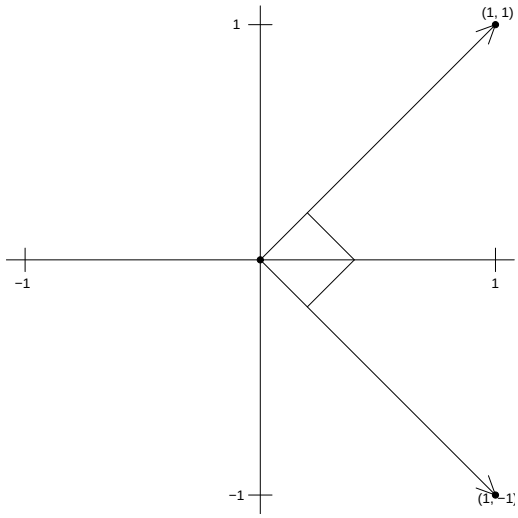


Figure 3-2: Illustration of two orthogonal contrasts: (1,1) and (1,-1).

You can check that every row is orthogonal to every other row.

This set of three row contrasts over four groups is also orthogonal

$$\begin{array}{rrrr} -3 & -1 & 1 & 3 \\ 1 & -1 & -1 & 1 \\ -1 & 3 & -3 & 1 \end{array}$$

The second set is called a set of polynomial contrasts. The first row is a "linear contrast," the second row is a "quadratic contrast," and the third row is a "cubic contrast." Many advanced textbooks print tables of polynomial contrasts for different number of groups (e.g., Hays; Kirk p. 814, Table E-10; Maxwell & Delaney, Table A.10, p A-25).

SS for contrasts

Contrasts partition SSB into smaller pieces that test more specific hypotheses. The reason for preferring orthogonal contrasts is because a complete set of orthogonal contrasts divides SSB perfectly (i.e., SSB = $SSC_1 + SSC_2 + \ldots + SSC_c$). Each SSC_i has one degree of freedom associated with it. Thus, if SSB has T - 1 degrees of freedom, we can find sets of T - 1 orthogonal contrasts that perfectly partition SSB. But different sets of orthogonal contrasts partition SSB in different ways. Further, the sum of squares for a

particular contrast is given by

$$SSC_i = \frac{\hat{l}^2}{\sum \frac{a_i^2}{n_i}} \quad (3\text{-}11)$$

See Appendix 1 for an example using the sleep deprivation data. One thing to notice is that the one-way ANOVA is simply combining the results from a set of T - 1 orthogonal contrasts. Any combined set of T - 1 orthogonal contrasts will yield the identical result for the SSB and for the omnibus test. There are many different sets of T - 1 orthogonal contrasts that one could use; you let your research hypotheses guide your selection of contrasts.

A simple pie chart can illustrate the decomposition of the sum of squares between groups (SSB), or sometimes called sum of squares treatment. Suppose that in an example we have SS treatments equal to 83.50. A set of orthogonal contrasts decomposes the portion of the pie related to SS treatment to smaller, non-overlapping pieces as depicted in the pie chart in Figure 3-3. The sum of squares within remains the same regardless of the choice of contrasts[3].

Why do we lose one contrast (i.e., T - 1)? One way of thinking about this is that the grand mean gets its own contrast, the unit contrast (all ones). Some statistical packages automatically give the test of significance for the unit contrast. It is a test of whether the average of all scores is different from zero. In most applications, the unit contrast carries little or no meaning. If a set of contrasts includes the unit contrast, then, in order for the set to be mutually orthogonal, all contrasts other than the unit contrast must sum to zero (i.e., $\sum a_i$). The restriction of having the weights sum to zero occurs so that contrasts can be orthogonal to the unit contrast. This guarantees that all contrasts are orthogonal to the grand mean. So, there are actually T contrasts to be had, with the grand mean μ taking one contrast and the αs taking the remaining T = 1 contrasts.

There is nothing wrong in testing sets of nonorthogonal contrasts, as long as the researcher is aware that the tests are redundant. Consider the test of all possible pairwise comparisons of the four groups in the sleep deprivation study (Appendix 2)—here we have six contrasts created from all possible ways of assigning a 1 to one group, a -1 to another group, and 0s to the rest. From the perspective of redundancy (i.e,. nonorthogonality) there is no problem here. However, from the perspective of "multiple tests" we may have a problem doing six tests at $\alpha = 0.05$.

(d) Corrections for multiple planned comparisons

[3]There are some approaches in the case where one is only interested in testing fewwer than T -1 contrasts that add the sum of squares for the untested contrasts ito the SSW. This practice has been hotly debated. I prefer keeping SSW as is regardless of whether or not I tested fewer contrasts than permitted by the design of the study

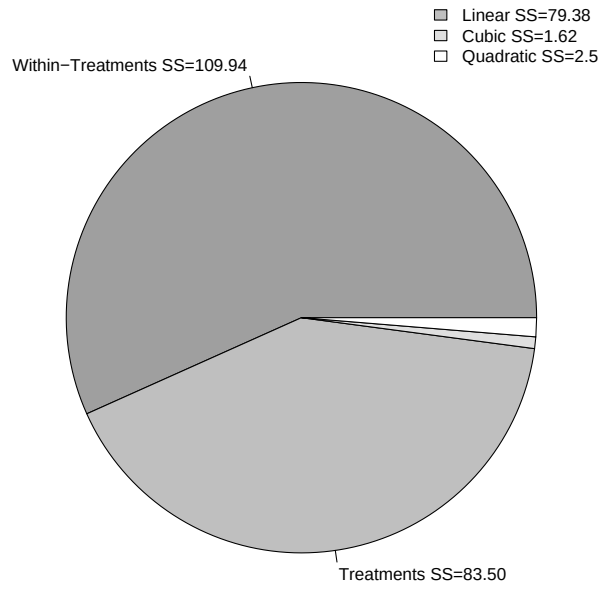


Figure 3-3: Pie chart illustrating decomposition of SS treatment into separate pieces by a set of orthogonal contrasts.

Contrasts give you flexibility to perform many different tests on your data. But, obviously there must be a limit to the number of comparisons one makes. If I perform 100 comparisons each at $\alpha = 0.05$, I would expect to find, on average, 5 significant differences just by chance alone assuming the null hypothesis is true in all 100 cases (i.e., I'd make a Type I error 5% of the time). Let's say that out of the 100 comparison I find 12 significant results—which ones are true and which ones are error?

There is a different question we could ask when someone conducts many tests. What are the chances of at least one Type I error some place across all the tests? The answer is that the rate grows quickly with the number of tests. Let me illustrate with a simple coin tossing example. If you toss a fair coin once, the chance of one head is .5. If you toss a fair coin twice, the chances of at least one head in two tosses is .75. In three tosses, the chances of at least one head is .875. The general formula for k tosses is

$$\text{P(at least one head in k tosses)} \quad = \quad 1 - (1 - p)^k \tag{3-12}$$

where p is the probability of the event occurring on a single trial (in the case of the fair coin, p = .5) and k is the number of trials.

We can use Equation 3-12 to answer the question of at least one Type I error in k tests. The probability of a Type I error for a single test is .05 (when the null is true), so we have for k tests

$$1 - (1 - \alpha)^k \tag{3-13}$$

Try this formula for different numbers of trials. When k = 5, the chances of at least one Type I error is .23. With 10 tests, the chances are .4, and with 25 tests the chances are .72. Thus, the more tests you perform, the more likely you are to have made at least one Type I error.

Dunn-Bonferroni: test individual contrast at a smaller α'

Some have argued that the Bonferroni correction should be used when making many comparisons. The Bonferroni correction finds a new α' to use at the individual contrast level that yields α as the family (or experiment-wise) error rate. The correction is

$$\alpha \quad = \quad 1 - (1 - \alpha')^{\mathrm{c}}, \tag{3-14}$$

where c is the number of contrasts being tested at α'. Note the similarity with Equation 3-13. FYI: the developer of this approach was Olive Dunn. She used the Bonferroni inequality in her proof to obtain a lower limit to the relevant probability. In keeping with tradition of how tests are named, this should be called the Dunn correction, or possibly the Dunn-Bonferroni correction.

So, if you want an overall α level of 0.05 when performing c comparisons, you can solve Equation 3-14 for α' and get

$$\alpha' \quad = \quad 1 - (1 - \alpha)^{\frac{1}{\mathrm{c}}} \tag{3-15}$$

Kirk presents a table (E-15) that gives *t* values corresponding to Equation 3-15 given α and c. The Bonferroni correction makes a test more conservative by reducing α to α'. Here α' is the new criterion you would use for each contrast.

A quick and (not so) dirty approximation to Equation 3-15 is

$$\frac{\alpha}{c} \approx \alpha' \tag{3-16}$$

So to apply this test all you do is evaluate observed p-values against α' instead of the usual α. If you perform 10 tests, then you would use the α criterion of 0.005 rather than the usual 0.05. [4]

Kirk presents a table (E-14) that gives *t* values corresponding to Equation 3-16 given α and c. Both of these procedures (i.e., Equation 3-15 and Equation 3-16) were proposed by Dunn and others. The former involves a "multiplicative inequality" and the latter involves an "additive inequality." The multiplicative inequality tends to work a little better (in terms of correction and power). If you don't have Kirk's table, you can use the table in Maxwell & Delaney, p A-10. It gives the Bonferroni adjusted F rather than t, and corresponds to Kirk's table E-14, the additive inequality (see Maxwell & Delaney, pages 202-, for a discussion).

These corrections apply to c **orthogonal** contrasts. If the contrasts are nonorthogonal, then the Bonferroni ends up overcorrecting (i.e., it's too conservative).

Anyone who makes 100 comparisons is probably on a "fishing expedition"; they deserve to have conservative criteria such as Bonferroni imposed on them. If they had clear, focused hypotheses they wouldn't need to do so many comparisons. Some fields like sociology and economics are forced to make many comparisons because of the nature of the domain they work in. Sometimes when we find ourselves performing 1000's of tests (as in the analysis of fMRI or genomics data) some people argue that a correction should be used. However, usually if one is performing thousands of tests, then probably the problem is not formulated correctly; it may be possible to reformulate the problem so fewer tests are performed.

I typically don't have the knee-jerk reaction "lots of tests means you automatically need to use Bonferroni." Most research questions asked by psychologists can usually be tested with a handful of comparisons. So, while I agree with the belief that those who make many comparisons should be punished and have heavy fines placed on them in the form of, say, Bonferroni corrections, in most applications we make a handful of planned contrasts and Bonferroni isn't a major issue.

Why hold this belief? Three reasons come to mind. Two are in the form of examples.

[4] This approximation results from taking the first order Taylor series approximation of Equation 3-15.

OK, I admit that I'm writing this late at night

i. Imagine two researchers, call them "Bonehead" and "Bright" (you can tell which side I take). Both are working on the same topic but in different labs. Bonehead devises a test of the question by comparing two groups. Later, he realizes that two additional groups must be run in order to have a better test of the hypothesis. Bonehead writes up his results as two different experiments each with a two sample *t* test. Bonehead receives praise for writing a wonderful paper with two studies. Bright, on the other hand, had the foresight to realize that four groups would be needed to test the hypothesis. So he runs all four groups, performs a oneway ANOVA with two contrasts. One contrast is 1, -1, 0, 0; the other contrast is 0, 0, 1, -1. So, conceptually both Bonehead and Bright are making the same two comparisons between pairs of means (there is a slight difference in the degrees of freedom between these two situations but we will ignore that here). I'm also assuming that the equality of variance assumption holds. The difference is that Bright, despite having the smarts to run all four groups from the outset, is asked to do a Bonferroni because he is making two comparisons, but Bonehead is given a pat on the back for running two experiments and a correction for two tests doesn't come to mind.

   One must demand that Bonehead and Bright apply the same criteria because both are performing identical tests. In my opinion, neither should be penalized. But if you believe a penalty should be applied, then it should be applied to both.

ii. The second reason will make more sense after we discuss factorial ANOVA, but I'll give it a shot now. Consider a 2×2 between subjects ANOVA. It yields three different tests: two main effects and one interaction. Each is tested at $\alpha = 0.05$. Even though most people don't notice this, the three tests face the identical multiple comparison problem as contrasts, yet journal editors don't complain about the multiple tests. Even more extreme, a three way ANOVA has seven comparisons each at $\alpha = 0.05$ (three main effects, three two way interactions, and one three way interaction).

   As we will see later, an equivalent way to test for main effects and interactions in a 2×2 ANOVA is to perform three contrasts. The contrast 1, 1, -1, -1 codes one main effect; the contrast 1, -1, 1, -1 codes the other main effect; the contrast 1, -1, -1, 1 codes the interaction. But, if I opted to test the 2×2 ANOVA through contrasts people would wave their arms saying "You're doing three contrasts. Bonferroni!" Recently, some applied statisticians have argued that main effects and interactions should be corrected by Bonferroni as well. Another example of a hotly debated issue. Personally, I think that is overkill and corrections aren't required in the case of factorial designs especially if each factor has two levels.

iii. There is a sense that the whole idea of family error rates and experiment-wise error rates is silly. Where does one stop correcting for error rates? Perhaps journal editors should demand article-wise error rates? How about correcting all tests in

a journal volume so the volume has a corrected α rate of 0.05? Or correct for department-wise error so that all tests coming out of a single Psychology department in a calendar year have a corrected Type I error rate of 0.05? How about career-wise correction?

The critical point is that we must be sympathetic to the idea that performing many, many comparisons is a bad thing with respect to Type I error rate. Performing all possible pair-wise tests in the context of a fishing expedition may be pushing the boundary on reasonableness of Type I error due to multiple comparisons. Not only do you have the problem of multiple comparisons when performing all possible pairwise tests, but you also suffer from the fact that there is a great deal of redundancy in those tests because the set of contrasts is nonorthogonal. We will deal with a better approach for handling the reasonable action of testing all possible pairwise tests when we discuss the Tukey test.

replication

If you are concerned about inflated Type I error rates due to multiple comparisons, then the best thing to do may not be a Bonferroni-type corrections but rather *replicate the study*. A replication is far more convincing than tinkering with α and is more in line with the scientific spirit. Don't play with probabilities. Replicate!

However, if you are doing an exploratory study with many, many comparisons, then you should make your α more conservative to correct for the inflated Type I error rate.

2. Assumptions in ANOVA

assumptions again

Contrasts have the same assumptions as the two sample *t* test (LN1) and between-subjects ANOVA (LN2):

(a) Independent samples (the ϵ_{ij}'s are not correlated)

(b) Equal population variances in each group (this allows pooling)

(c) Normal distributions (the ϵ_{ij} within each group are normally distributed)

3. Checking assumptions.

(a) Be careful when using statistical tests of variances such as the Hartley test or Box's improvement of Bartlett's test (Bartlett-Box) or Cochran's C. All three are very, very

sensitive to departures from normality. I mention these tests because they are in SPSS. There are better tests than these (such as Levene's test, which is based on absolute differences rather than squared deviations, and O'Brien's test), but my advice is to avoid them all together. If you violate the equality of variance assumption, you'll know it (the variances will obviously be different). If you look at the boxplot before analyzing your data, you'll catch violations.

There is a peculiar logic in using a hypothesis test to test an assumption. The hypothesis test used to test the assumption itself makes assumptions so you are in a funny position that you'll need to test the assumptions of the test on the assumptions you are are testing, and so on. Also, as you increase sample size, the power of the hypothesis test increases. So, with a large enough sample you will always find a violation of the assumption no matter how tiny the ratio between the smallest observed variance and the largest observed variance.

The statistician Box once said "To make a preliminary test on variances is rather like putting to sea in a rowing boat to find out whether conditions are sufficiently calm for an ocean liner to leave port" (1953, *Biometrika, 40*, p. 333).

(b) You could use the trusty spread and level plot I introduced in the previous lecture notes to help you decide what transformation to use

nonparametric alternative

(c) Kruskal-Wallis test as an option when the normality or equality of variance assumptions are violated. The test is designed for normality violations but because it is identical to an ANOVA on ranks it inadvertently can correct some violations of variances.

The Kruskal-Wallis test is a simple generalization of the Mann-Whitney U (analogous to the way ANOVA generalizes the two sample *t* test). The Kruskal-Wallis test is equivalent to performing an ANOVA on data that have been converted to ranks[5]. Kruskal-Wallis yields an omnibus test. Some nonparametric procedures for testing post hoc comparisons exist. In this context planned comparisons are simply the usual contrasts defined on the ranked data (where the error term comes from the ANOVA on the ranked data). See Appendix 3 for an example of the K-W test using the sleep deprivation data.

The Kruskal-Wallis test has not been generalized to complex, factorial designs, so one cannot go very far when interactions need to be tested (actually, some generalizations to the two-way ANOVA have been proposed but they are not on strong theoretical footing). About the only way to deal with violations of assumptions in complicated designs is to perform transformations.

[5]For a mathematical proof of this see Conover, *Practical Nonparametric Statistics*.

For a review of developments in nonparametric tests see Erceg-Hurn & Mirosevich (2008), *American Psychologist, 63,* 591-.

An aside: Also, some nonparametric tests, such as the Mann-Whitney U (a special case of of the Kruskal-Wallis test), have special interpretations that one can use to make inferences in some pretty nasty situations, such as when independence is violated. Sometimes even in a randomized experiment one violates the independence assumption because one case in the treatment condition may influence others. For example, a vaccine may be effective at reducing susceptibility to a disease, but one person not getting the disease because they received the vaccine may in turn reduce the chances of another person (say in the control condition) getting the disease because now there is less opportunity for contagion. Another example: if a half the kids on a class get one treatment, and that makes the teacher treat all kids in the class differently, then there is interference in that one unit being assigned a treatment influences other units. It turns out that some nonparametric statistics like the Mann-Whitney have a special interpretation that when used in a special way make them robust to such violations of independence. The M-W can be interpreted as the probability that a data point in the treatment condition will exceed a data point in the control condition. The null distribution (no effect) for this test does not depend on the observed data but merely on the number of subjects in each of the two conditions, so independence is not an issue for the null distribution. By comparing two M-W tests (one on the observed data and one on the null data), one can get around these types of independence violations. An elegant paper by Rosenbaum (2007, JASA, 102, 191-200) develops this point in a rigorous way, along with extensions to covariates. It is interesting that the early days of experiments used randomization tests like the Mann-Whitney, but that became too difficult to do by hand with complicated experimental designs, so people like Fisher developed shortcuts such as ANOVA. The shortcuts involved additional assumptions like normal distributions, equal variances, and independence. Now with modern computational power we are seeing a return to distribution-free tests such as randomization tests that don't make those assumptions. In a decade or two people may view ANOVA and other techniques that make "unnecessary assumptions" such as normality, equality of variances and independence as quaint tools.

(d) Another option is to perform the Welch test on the contrast so that you aren't making the equality of variance assumption when testing the contrast. But the Welch computation for pooled variances doesn't exist for omnibus tests, so it won't help if someone requires you to perform the omnibus test. Further, the Welch correction will only help with the violation of equality of variances; it won't help with other violations such as violations of normality.

4. The meaning of omnibus significance tests

I have done my share of criticizing the omnibus test. Expect more of that behavior as we move into complicated ANOVAs where omnibus tests occur almost at every turn. Throughout

I will show how omnibus tests are related to individual pieces such as pairwise comparisons and contrasts.

In a homework problem I'll have you verify that the omnibus test is equivalent to the average of t^2s from all possible pairwise contrasts. This result implies that if the omnibus F is significant, then at least one of the pairwise contrasts is significant. This is one justification for Fisher's suggestion that the omnibus test be conducted first to see if the "light is green" with respect to performing all possible pairwise comparisons. We will discuss better strategies that solve the multiplicity problem directly (e.g., Tukey's test).

Can a contrast be significant even though none of the pairwise differences are significant? Yes, but in that case the omnibus test will not be significant. Can you explain why?

It turns out that if you take the average F (or t^2s) from an orthogonal set of contrasts, that will also equal the omnibus F. This is easy to verify. The hint is to use the equation F = SSC/MSE, which is based on the property that contrasts always have one degree of freedom in the numerator. The average of Fs over T - 1 orthogonal contrasts will yield MSB/MSE, the omnibus F. Be careful if you try to verify this on an example with unequal sample sizes as orthogonal contrasts need to take into account sample size as well.

So, if the omnibus test is statistically significant, then at least one contrast is significant. But it might be that the omnibus test is not significant when at least one of the contrasts is. This can happen whenever the sums of squares for a particular contrast (SSC) is greater than SSB/df (sums of squares between groups divided by the degrees of freedom for between). Recall that SSB is itself a sum of sum of squares of an orthogonal set of contrasts, meaning SSB=SSC_1 + SSC_2 + . . . + $SSC_{(T-1)}$ for T - 1 contrasts.

Summary

The omnibus F is equal to an average of orthogonal contrasts (which is based on T - 1 comparisons) and it is also equal to an average of pairwise comparisons (which is based on T(T - 1)/2 comparisons). Both averages give the same result but are based on different numbers of comparisons. They provide different but equivalent ways to interpret the omnibus test.

5. Power and Contrasts

I mentioned in an earlier lecture notes that computing power is typically not easy. If you want to understand power and its computation at a deep level, look at the Cohen book that is mentioned in the syllabus.

I recently worked out some approximations that seem to be close enough for many applications involving contrasts. The following works only for F tests that have one degree of freedom in the numerator. All contrasts have one degree of freedom in the numerator so this

approximation works for contrasts (it will also work for parameters like correlations, beta coefficients in regressions, and any ANOVA main effect or interaction where each factor has two levels such as in a $2 \times 2 \times 2$ design). I have not studied the boundary conditions of these approximations carefully so use them at your own risk.

The convention is to strive for a power of at least .80, which is another way of saying that you are setting the Type II error rate to be at most .20 (i.e., 1 - power).

One needs to be careful when computing power from observed estimates. The estimates from a sample follow a distribution and you need to keep in mind that if one were to redo the study the estimates would be slightly different due to sampling and error, and hence one would get different power estimates.

- Case I: there is a result in the literature that you want to replicate and you want to figure out how many subjects you need in order to achieve adequate power. Do you need more subjects than the original study? Can you get by with fewer subjects? Do you need the same number of subjects as the original study? You'll need to know the original sample size N_1 of the study you are going to replicate and the F value (or if a t is reported for the contrast, just square the t).

$$[\text{new sample size required for power = .8}] \approx \frac{9N_1}{F_{\text{obs}}} \tag{3-17}$$

  For example, if the observed F in the first study was 9 (I'm choosing 9 for the observed F in this example so that it equals the constant and they cancel and the required sample size for .80 power is N_1), then the same sample size as the original study will give you a power of approximately .8. In other words, you need relatively large effects (as indexed by an observed F of at least 9) in order to use the same sample size in a replication and have conventional levels of statistical power. In terms of p-values, an observed F of 9 corresponds to an observed p-value of roughly .005 (exact values depends on degrees of freedom), so if you want to replicate a study that had an observed F of 9 (or equivalently a p value of roughly .005) you can achieve a power of roughly .80 using the same sample size as the original study. This is one rationale for why some people argue for using α = .005 as a criterion because that tells me the observed study has at least an 80% chance of replicating with the same sample size.

  By comparison a p-value of .05 corresponds to roughly an observed F of 4 (exact value depends on degrees of freedom) so that would require a sample size in the replication more than double (i.e., 9/4 = 2.25) the original sample size to achieve at least an 80%

chance to replicate.[6]

The 9 in the approximation comes from another approximation to the noncentral F distribution under the conditions specified here relative to the choices for the desired power of .80 and a two-tailed α level of .05. The constant is actually 8 and change, so I'm rounding up to provide a conservative estimate to the required sample size. The size of the "change" depends on the degrees of freedom (a combination of N_1 and number of groups). When sample size is small at, say, 10, the constant is 8.67; as sample size N_1 increases, the constant approaches 8.07 in the limit as N_1 approaches infinity. That is why I just took 9 as a conservative approximation to the constant so that we don't have to worry about these little details.

- Case II: you are doing a brand new study and want to have sample sizes large enough to detect a particular effect size. This is the classic case of computing power because it does not depend on observed estimates. First, choose a measure of effect size; there are many possible choices and they each have different properties. For a contrast you can use this effect size measure: estimate the proportion of sum of squares contrast (SSC/SST) and the proportion of the between sum of squares (SSB/SST). That is, estimate what proportion of the variability you predict/hope your contrast will account for (call this p_c = SSC/SST) and estimate the total proportion of variability accounted for in your study (p_b = SSB/SST). Note that p_b must be greater than or equal to p_c (think of the pie chart). Finally, to estimate sample size plug into this approximation:

$$[\text{sample size required for power = .8}] \quad \approx \quad \frac{9(1-p_b)}{p_c} \tag{3-18}$$

  You can use this Case II framework, for example, in cases where you may not know the effect size but have a guess of the minimum effect size that would be useful (that is, any effect size below that value would not be of interest). Here, rather than guess an effect size you are drawing a line in the sand and saying I want to power my study (i.e., have sufficient sample size) to detect effect sizes of at least that size.

  The 9 in this formula carries the same meaning as in Case I.

- Case III: you ran a study, failed to find a significant result but you don't want to give up. You think that you need more power so you decide to run more subjects. How many more subjects should you run? Just follow the procedure outlined in Case I, using the sample size and F observed from the study that failed to find the result.[7]

[6] These are rough approximations to power in the replication sample. In reality, this is a more complicated problem because there is also publication bias, issues of correcting for multiple tests, other sources of noise in addition to the usual SSW such as measurement error, variability due to selection of stimuli that may not have been properly accounted for in the computation of SSW (for this type of deeper analysis to power, random effect models may be a good option, we'll cover those in LN4 and LN9), etc.

[7] I'm not advocating repeatedly checking if a few more subjects would get one over the .05 hump. That would be wrong. If your first set of subjects produced an F of, say, 3, then Case I would say run 3 times the number of subjects you have run so far. That is much more than just run a few more and see what happens. Other philosophical views in statistics, such as the Bayesian view, have different approaches to stopping rules when collecting data. In some Bayesian frameworks, it is perfectly fine to make repeated decisions to collect more data because the framework models updated belief up to the data collected so far. But unless we completely adopt the Bayesian perspective we are limited by the machinery of the frequentist approach.

Be careful about putting too much stock into the observed power, which is a function of the p-value (see Hoenig & Hiesey, American Statistician, 2001, or Greenwald et al, 1996). Some programs like SPSS provide observed power estimates; I recommend you do not use those estimates.

G*Power

There are useful programs becoming available for computing power in many different situations, though admittedly most are very simple situations. G*Power is a commonly used open-source and free program available on all three major platforms (PC, Mac and Linux). It has a nice point-and-click interface to power. The URL is http://www.gpower.hhu.de/.

For contrasts in G*Power select F-test and linear multiple regression ... increase. We will focus on power for testing orthogonal contrasts and ignore the power for the omnibus F test. To do an example, click on determine effect size f, and enter 15 for effect variance and 50 for residual variance, then click on calculate and transfer to main window. Here I'm assuming that total variance is 100, the SSC for the contrast of interest is 15 and the SSW is 50 (so SSB is 50 and other two contrasts have a combined SS of 35). Then, set α to .05 and power to .80, one test predictor and 3 total predictors (this would correspond to a total of 3 contrasts, so that implies 4 groups, and we want to know the power for one of the contrasts). The computation yields a needed sample size of 29 and an estimated power of .81. To compare this to the approximation I presented earlier, using the assumptions I made above we have p_b = .5 and p_c = .15. Plug that into my formula and needed total sample size is 30, which is pretty close to the G*Power estimated sample size. Within G*Power the partial R^2 is SSC/(SSW+SSC), or in this example 15/65, and f^2 is partial R^2/(1- partial R^2), which for this example is .2308/(1-.2308) = .3.

If you want to learn the "real" way to compute power analysis (rather than the approximations I present here) you can look at some of the recommended textbooks in the syllabus. The discussion of power in KNNL is a little easier to follow than the discussion in MD. Table B5 in KNNL can be used for any F test with one degree of freedom in the numerator, hence it can be used for contrasts too. Indeed, looking at column for δ = 3 page 1346 shows that the power for reasonable samples sizes is .80 or thereabouts, which is consistent with the approximation I gave above because $\delta^2 = F$. I will not ask you to compute power on exams.

Be careful about just using effect size measures because someone else used them. Some effect size measures lose information about direction of effect (like variance accounted for measures), which can screw up a meta-analysis. There are formulas to convert effect size measures into different effect size measures, so you may need to convert an effect size you see in the literature into something else to help you interpret and compute power.

SPSS power

SPSS computes power within the MANOVA command. We'll learn the ins and outs of the MANOVA command later, but I'll give the syntax of the subcommands here for reference. All you have to do is add these two lines to your MANOVA syntax and you'll get power

analyses in the output. Unfortunately, the output won't tell you what sample size you need but you can use the output and plug into the CASE I approximation above. This SPSS feature won't help you with CASE II because SPSS assumes you already have data in hand. More on MANOVA later.

```
/power exact
/print signif(all) parameters(all)
```

R power

R has a few features for power analyses such as the power.t.test() command in the pwr package. There are a couple packages in R for doing power (pwr and poweR) that go into much more detail and require a little startup time to learn how to use. There are some power packages with Shiny interfaces (a graphical user interface to permit point-and-click use). There are also some packages that provide utility functions to compute power through simulation such as paramtest, simr and others. These packages require quite a bit of programming to set up the model structure you want to test and the simulation to compute power and sample size. For example, one needs to cycle through different values of possible effect sizes, for each of those possible values create, say, a 1000 data sets, compute ANOVAs/contrasts on each data set and then tabulate or plot the summary analyses (e.g., for a given sample size and a given effect size, what proportion of the simulated 1000 data sets correctly rejected the null hypothesis). You're searching for the sample size that in combination with an effect size leads to a desired power level such as .80. I won't go over this in class. If you are interested in learning more about simulations to estimate power and sample size requirements and seeing worked out examples, take a look at the documentation and vignettes for these packages.

Here is an illustration in R using the pwr package with the same four group example I used to illustrate G*Power. It is a little different than the example in G*Power because it isn't the isolated contrast test accounting for the other orthogonal contrasts. We enter the numerator degrees of freedoom (1 for a contrast), the effect size in the f^2 metric, the significance level and desired power. The function returns the denominator degrees of freedom so we have to add back the number of groups to get sample size (i.e., it reports df = N –T so to estimate the needed N we compute df + T, where T is the number of groups).

```
library(pwr)
pwr.f2.test(u = 1, f2 = 0.3, sig.level = 0.05, power = 0.8)
```

```
##
##      Multiple regression power calculation
##
##               u = 1
##               v = 26.2189
##              f2 = 0.3
```

```
##        sig.level = 0.05
##            power = 0.8
```

The degrees of freedom for denominator is 26 so add 4 for the number of groups and the result is 30 subjects total, so rounding up 8 subjects in each of the 4 groups. Again, this is a little different from G*Power (in this case G*Power yielded 29) because it did not account for the other two contrasts like we did in G*Power but for this example we get the same sample size estimate as the approximation I offer. A partial f^2 is considered a large effect so that is why the sample size comes out to be relatively small. Effect sizes more in line with what are seen in the social sciences are much smaller and can be in the vicinity of $f^2 = .05$, or even smaller. Plugging that effect size into the function yields a sample size of 160, or 40 per group, which is quite a big jump from 8 per group.

```
pwr.f2.test(u = 1, f2 = 0.05, sig.level = 0.05, power = 0.8)
```

```
##
##      Multiple regression power calculation
##
##               u = 1
##               v = 156.9209
##              f2 = 0.05
##       sig.level = 0.05
##           power = 0.8
```

A good heuristic to follow is to ask what kind of effect sizes you want to be able to find if they are present and power your study accordingly. Follow the analogy of the fishing net. By deciding on the minimum size of the fish you want to catch, and are happy ignoring the smaller fish, you can then select the appropriate net to use. If you select the minimum effect size you want to be able to detect in your study, power analysis provides a method for selecting the appropriate sample size (the fishing net with the appropriate mesh) so you have a reasonable chance (such as 80% power) of detecting effects of that size.

6. *Post hoc* comparisons

There is an important distinction between planned and post hoc comparisons. If you are performing planned contrasts, then there is an argument that worrying about Type I error corrections may not be needed. But, if you are performing comparisons on the basis of having seen the results ("Oh, look! The mean for Group A is greater than the mean for Group B. I wonder if it is significant?"), then post hoc tests are appropriate, which perform an adjustment to maintain the Type I error rate.

You might think, "Why not just do a Bonferroni correction on *post hoc* comparisons?" Consider, for example, the sleep deprivation study with four groups. Suppose you wanted to test all possible pairwise comparisons between the four means. Why not divide the desired overall α by six (with four groups there are six possible pairwise comparisons)? There isn't anything fundamentally wrong with doing that, but there are two limitations to this "solution." First, the Bonferroni correction applies to orthogonal contrasts, but the set of all possible pairwise contrasts is not orthogonal. For nonorthogonal contrasts the Bonferroni correction is conservative (i.e., it overcorrects). Second, with a large number of groups the Bonferroni correction is a tough criterion. Even with four groups, the per comparison α' needed to achieve an overall $\alpha = 0.05$ is 0.008 because there are six possible pairwise comparisons. Fortunately, there is a better procedure for testing all possible pairwise comparisons.

7. The problem of deciding to test a difference after snooping the data

Consider this situation (from Maxwell and Delaney, 1990). A researcher wishes to test four means. Among other things, she planned to test whether $\mu_2 - \mu_4 = 0$. Now compare that to the situation where the researcher had no specific predictions about the four means. Instead, after looking at the data she observes that $\overline{Y}_2$ is the maximum of the four means and $\overline{Y}_4$ is the minimum. She wants to test whether the maximum mean is significantly different from the minimum mean. But that is actually a test of

$$\mu_{\text{max}} - \mu_{\text{min}} = 0 \tag{3-19}$$

The difference in these two cases (planned and post hoc) is that we know the probability of making a Type I error in the case of planned contrasts. However, the sampling distribution of $\overline{Y}_{\text{max}} - \overline{Y}_{\text{min}}$ (the difference between the maximum and the minimum mean) is very different, making P(Type I error) $\geq \alpha'$.

The sampling distribution of the "range" (i.e., max - min) is part of a more general family of order distributions. We will spend time developing intuition on order distributions using an example adapted from Winer (1971). Consider a population having the following five elements

Element	X
1	10
2	15
3	20
4	25
5	30

Draw samples of size 3 (without replacement). What does the sampling distribution of the range look like? How about the sampling distribution of the difference between the middle

score and the minimum score? With this small population these two sampling distributions are easy to create. Simply list all possible samples of size 3,

sample	d_2	d_3
10, 15, 20	5	10
10, 15, 25	5	15
10, 15, 30	5	20
10, 20, 25	10	15
10, 20, 30	10	20
10, 25, 30	15	20
15, 20, 25	5	10
15, 20, 30	5	15
15, 25, 30	10	15
20, 25, 30	5	10

where d_2 and d_3 are defined as 1) the difference between the middle score and the minimum score and 2) the difference between the maximum score and the minimum score (i.e., the range), respectively.

The frequencies of each value of d_2 and d_3 are

d_2	frequency	d_3	frequency
20	0	20	3
15	1	15	4
10	3	10	3
5	6	5	0

If each of these samples are equally likely, then, obviously, the probability of large values for d_3 must be greater than the probability of large values for d_2.

In sum, when you snoop and test differences between, say, the minimum and maximum means in a study, you are actually doing a test on the range and must take into account the sampling distribution of the range, not the sampling distribution of the mean. The tests presented below incorporate the sampling distribution of the range, which is not the same as the sampling distribution of the mean and doesn't follow the central limit theorem.

8. Statisticians have studied the relevant sampling distributions involved in "snooping" and have developed tests that make the proper adjustments. Here is a brief discussion of a few post hoc procedures. I will focus on what the tests do, what they mean, and how to interpret the results. Examples will be given using the sleep deprivation data.

(a) Tukey's W procedure (aka "Honestly Significant Difference" test)

Tukey grasped the problem of the distribution of $\overline{Y}_{max} - \overline{Y}_{min}$ and developed an appropriate test. He showed that the standardized range follows a special distribution (the tilde character refers to "the right hand side follows the distribution on the left hand side"):

$$q_\alpha(T,v) \sim \frac{\overline{Y}_{max} - \overline{Y}_{min}}{s_{pooled}\sqrt{\frac{1}{n}}}$$

where T is the number of groups, v is the degrees of freedom associated with MSE, q_α(T,v) is the studentized range distribution (a table lookup), $s_{pooled} = \sqrt{MSE}$ and n refers to cell sample size. The studentized range distribution q_α(T,v) is rather complicated. It has two types of degrees of freedom. One refers to the number of groups (T), the other refers to the degrees of freedom associated with MSE (as in the one way ANOVA, v = N - T).

The studentized range distribution is presented in table form in Maxwell and Delaney Table A4. An excerpt from the table for 4 groups with 28 degrees of freedom corresponding to the error term (N - T) for "family wise error rate" equal to 0.05. The table doesn't list $df_{error} = 28$ so for this example I'll use the next closest but smaller value of 24 (that's why I put an approximation sign next to the table entries for $df_{error} = 28$); I won't bother with interpolation issues here.

Tukey critical values in R

You can get these critical values for Tukey using the R command qtukey(). Entering qtukey(.95,4,28) produces as a value of 3.86.

Excerpt from the studentized range table

df_{error}	$\cdots$	$T = 4$	$\cdots$	$T = 20$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
28	$\cdots$	≈ 3.90	$\cdots$	≈ 5.59
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
∞	$\cdots$	3.63	$\cdots$	5.01

The test is relatively simple. You compute a critical difference W that the difference between a pair of means must exceed,

$$W = q_\alpha(T,v)\sqrt{\frac{MSE}{n}} \quad (3\text{-}20)$$

The MSE term comes straight out of the ANOVA source table. If an observed absolute difference between two means, $|\overline{Y}_i - \overline{Y}_j|$, exceeds W, then the null hypothesis that $\mu_i - \mu_j = 0$ is rejected. W plays an analogous role to the "chains" in football to mark first down.[8]

Tukey's W allows you to test all possible pairwise means while maintaining an overall Type I error rate of α. This is because everything is calibrated to the range distribution between the max mean and the min mean, so any other pair of means in the study will have a difference smaller than the range.

Again, the outline of the procedure is as follows. First, calculate the minimum difference between means that is needed to achieve significant results (i.e., W). Second, order your means from min to max. Finally, compare each observed difference between the means against W, the critical difference required for significance. If an observed difference exceeds the critical difference, then you reject the null hypothesis that $\mu_i - \mu_j = 0$. The null hypothesis is the same as the two sample *t* test, but we test it against a different distribution that takes into account that we are conducting tests on all possible pairs of means.

One could also construct confidence intervals around the pairwise differences (equal sample size formula)

$$\overline{Y}_i - \overline{Y}_j \pm q_\alpha(T,v)\sqrt{\frac{MSE}{n}} \quad (3\text{-}22)$$

$$\pm W \quad (3\text{-}23)$$

Computing pairwise comparisons in this format controls the Type I error rate at α across all confidence intervals in the set of pairwise comparisons.

Tukey is the method of choice; subsequent work shows this is the best method for controlling the Type I error rate in pairwise post hoc tests.

SPSS calls this test "TUKEY". In R this Tukey test is computed through the TukeyHSD() command.

(b) Newman-Keuls

[8] For unequal samples size use this formula instead

$$W_{ij} = q_\alpha(T,v)\sqrt{\frac{MSE}{2}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)} \quad (3\text{-}21)$$

Even though the Tukey test is the method of choice, psychologists frequently use a variant of Tukey's W: the Newman-Keuls test (SNK). As you will see, both tests are very similar. But, SNK is more liberal so it is more likely to reject the null hypothesis (maybe this accounts for why people use SNK more often than Tukey). SNK takes the approach that the farther apart two means are on an ordered scale, the larger the difference between them must be for the range to exceed its critical value.

The only difference between Tukey's W and SNK is in the value you look up in the studentized range distribution. While Tukey has you use the greatest value of q (i.e., given the T groups), SNK varies q depending on the number of means that are in between the two particular means you are testing. The SNK uses q_α(R,v) rather than q_α(T,v), where R is the number of means, or steps, between the two means that you are testing. Thus, the critical value W given by SNK will vary depending on the number of means between a given pair. However, this makes the precise overall Type I error rate of the SNK test somewhat ambiguous. Simultaneous confidence intervals for the SNK test cannot be constructed.

Tukey developed a compromise between his "honestly significant difference" test and SNK. One simply averages the q values from HSD and SNK and proceeds as usual, using the average q

$$\frac{q_\alpha(\text{T,v}) + q_\alpha(\text{R,v})}{2}$$

This is known as Tukey's B procedure (aka "wholly significant difference"). I don't recommend Tukey's B because the error rate it is controlling is ambiguous, but mention it because SPSS has it as an option.

(c) Scheffe

The Scheffe test allows you to test any contrast, as many as you want without regard to orthogonality. Intuitively, the test acts as though you are testing an infinite number of contrasts. This makes the critical value necessary to reach significance very strict. Consequently, the Scheffe test is one of the most conservative of the post hoc tests. It is one you should have in your bag of tricks because it is the way to test complex contrasts (i.e., contrasts more complicated than pairwise) that were not planned.

The contrast value $\hat{\text{I}}$ is compared against S, where

$$\text{S} = \sqrt{\text{V}(\hat{\text{I}})}\sqrt{(\text{T-1})\text{F}_{\alpha,\,\text{df1},\,\text{df2}}} \tag{3-24}$$

If $|\hat{\text{I}}| > \text{S}$, then you reject the null hypothesis that $\text{I} = 0$. Recall that

$$\text{V}(\hat{\text{I}}) = \text{MSE}\sum \frac{\text{a}_\text{i}^2}{\text{n}_\text{i}}$$

The square root of $V(\hat{I})$ is the standard error of $\hat{I}$ and $F_{\alpha,\ df1,\ df2}$ is the appropriate value from the F table, with df1 corresponding to the df of the omnibus test (i.e., T - 1) and df2 corresponding to the df for MSE (i.e., N - T).

The way Scheffe works is to consider the sampling distribution of $F_{maximum}$. The maximum F corresponds to the contrast that gives the greatest SSC (see Maxwell and Delaney, p. 215). If the omnibus test is significant, then there is at least one contrast that Scheffe will find to be significant. If the omnibus test is not significant, then Scheffe will not find any significant contrast.

To build a confidence interval around $\hat{I}$ simply add and subtract S to get the endpoints of the interval; that is,

$$\hat{I} - S \quad \text{and} \quad \hat{I} + S$$

SPSS shortcut

The Scheffe command implemented in SPSS is rather limited. It tests all possible pairwise comparisons; it does not allow you to specify arbitrary contrasts. But, once you have the *t* tests corresponding to the contrasts of interest from the ONEWAY output, the Scheffe test is easy to compute by hand. Recall that the *t* for the contrast is equal to $\frac{\hat{I}}{\sqrt{V(\hat{I})}}$. Therefore, Equation 3-24 can be re-expressed as testing the observed *t* against

$$\text{Scheffe } t_{critical} = \sqrt{(T\text{-}1)F_{\alpha,\ df1,\ df2}}$$

where the F is just a table lookup using degrees of freedom corresponding to the omnibus test (both for numerator and denominator)[9]. This means you can use SPSS output from the SPSS ONEWAY command to find Scheffe for any contrast, even though the built-in SPSS Scheffe doesn't permit anything other than pairwise comparisons. All you do is compare the observed t to the new t critical from Scheffe—you ignore the p-value printed in the output. If the observed t exceeds in absolute value sense the Scheffe t critical, then you reject the null hypothesis according to the Scheffe criterion.

A Welch version of the Scheffe test is easy to implement (discussed in Maxwell and Delaney). Compute everything the same way except use the degrees of freedom corresponding to the separate variance contrast for the df2 term in Equation 3-24.

(d) Duncan

Duncan's test is almost identical to the SNK test but Duncan uses a different distribution than the studentized range. Rather than q_α(R,v), the Duncan test uses q'_α(R,v). The main change in the q' is that the α levels have been adjusted. Actually, the theoretical

[9] Technically, this is not a t distribution because in general when you use the Scheffe the degrees of freedom numerator will be greater than 1, but the relation that $F = t^2$ only refers to the case when df numerator equals 1. I call this the Scheffe t critical value somewhat loosely and do not mean to imply it is a t distribution.

foundation of the Duncan test is quite interesting because it involves an empirical Bayes solution. However, it is not clear how useful the test is in actual data analysis. Two limitations of the Duncan test are 1) it is not clear what error rate is being controlled and 2) it doesn't lend itself to building confidence intervals around pairwise differences.

(e) Fisher's least significant difference (LSD) test

Don't worry about Fisher's LSD test. Fisher's LSD is an approximate solution that was developed before Tukey published the HSD test. LSD attempts to control the Type I error rate by establishing the decision rule that you can only do post hoc tests if you find a significant effect in the ANOVA. Unfortunately, this logic does not always apply (i.e., it is easy to construct counterexamples). It is this (fallacious) reasoning that led psychologist to think that you are allowed to perform post hoc tests only when you find a significant omnibus result in your ANOVA. Note that this rule of requiring a significant omnibus ANOVA before getting the green light to perform post hoc tests only applies when performing Fisher's LSD and not the other post hoc tests we consider in this course[10].

(f) Misc. tests

There are many other tests that we will not discuss in detail (see Kirk's *Experimental Design* for a detailed discussion of additional tests). I'll just mention some of the tests so you become familiar with the names. Dunn developed a multiple comparison procedure to test nonorthogonal, planned contrasts. This procedure was later improved upon by Šidák. In some situations, though, these tests give results that are very similar to just doing a Bonferroni correction. Additional post hoc tests include variations of the Tukey HSD test by people such as Brown & Forsythe and Spztøvoll & Stoline. Dunnett developed a post hoc test to compare a control group to several treatments. The list goes on.

false discover rate (FDR)

Another interesting measure is called the *false discovery rate* (also known as the Benjamini-Hochberg correction). It is becoming popular in some circles, such as fMRI research. FDR corresponds to the proportion of statistically rejected tests that are falsely rejected. Ideally, if a procedure behaves properly, the FDR should approach the α criterion. I won't go into the derivation or logic for this procedure, but the end result for *independent* tests is a slight modification of Bonferroni. Recall that in Bonferroni, you divide the overall α by the number of tests c, as in $\frac{\alpha}{c}$. The modified procedure under the FDR also considers the number of those tests that are statistically significant. Let's call that s. So if you perform $10 = c$ tests and $5 = s$ of those tests are significant, and you want

[10] Under some very special conditions such as a one-way ANOVA with three groups, Fisher's LSD test might be okay in the sense that it performs as well as (but never better than) Tukey's test (Levin et al., 1994, *Psychological Bulletin, 115,* 153-159). Levin et al discuss more complicated situations where there are no alternatives to Fisher's "check the omnibus test first" procedure.

$\alpha = .05$, you apply this formula:

$$\text{FDR} = \frac{s\alpha}{c} \tag{3-25}$$

But there is a slight deviation from simply applying this formula. One first orders the c observed p values in increasing order and select the first k tests for which the observed p value is less than or equal to $\frac{k\alpha}{c}$. For example, suppose you run 4 tests and all four are statistically significant by themselves. This means that c=4 and s=4. You then order the observed p-values in increasing order. Let's say the four observed p-values were .001, .0126, .04, and .045. In this example all four are each statistically significant by themselves. However, using the FDR we would only consider the first two as statistically significant by this procedure because the first two are below $\frac{1*.05}{4} = .0125$ and $\frac{2*.05}{4} = .025$, respectively. However the next one exceeds the corresponding values of this criterion $\frac{3*.05}{4} = .0375$ (note how the criterion changes for k=1, k=2, k=3, etc). The procedure stops there as soon as the next p-value in the rank order fails to exceed the criterion. Thus, even though the four tests have observed p-values less than .05, under this FDR criterion only the two smallest p-values are statistically significant; the other two are not because they exceed the FDR criterion. For comparison under the traditional Bonferroni only the first test exceeds the Bonferroni criterion of .05/4=.0125 so only one test would be significant in this example by Bonferroni. For a technical treatment of this procedure see Benjamini & Yekutieli (2001) and Storey (2002).

An entire course could easily be devoted to the problem of multiple comparisons. For a technical reference see the book by Hochberg & Tamhane, 1987, *Multiple Comparison Procedures*. A readable introduction to the problem of multiple comparisons appears in Larry Toothaker's book *Multiple Comparisons for Researchers*.

(g) Which post hoc procedure should one use?

It depends. What error rate is the researcher trying to control? Is the researcher performing pairwise comparisons or complex contrasts?

It is best to make an analogy with driving behavior and speed limits. Speed limits are clearly posted and they mark the law. In practice, people speed. Different people speed different amounts (e.g., 2-5 miles over, 10 miles over the limit)? Some people offer rationales like "I'm moving with the flow and it would be unsafe to go slower" or "I won't be pulled over for just a minor difference" or "I'm speeding now because I'm late for an important meeting or for picking up my child from daycare." The point is that there is the law and various behaviors people do that violate the law, and we provide various rationales to justify that behavior.

The situation of multiple comparisons is comparable to that. The basic principle (the "law") is that Type I error is inflated when we conduct many tests in the sense that the

probability of at least one Type I error increases the more tests we conduct. Now in different contexts researchers raise various rationales to justify their behavior to bypass the law. I planned my tests so I don't have to worry about Type I error inflation. I'm only doing these subset of tests, so I'm ok. I'm only doing pairwise tests, so I don't have to worry about Type I error inflation. These sound similar to the rationalizations we make to justify our driving. A lot of the advice that is out is in this spirit: making various exceptions or justifications for addressing this issue. And you can view the flow chart below in that spirit too.

Simulation studies (e.g., Petrinovich & Hardyck, 1969) suggest that if one is controlling for experimentwise error and doing pairwise comparisons, then Tukey's W is preferred because it has good power and an accurate Type I error rate. If one is doing arbitrary post hoc contrasts and wants to control experimentwise error, then Scheffe is the way to go. Finally, if you are testing a few orthogonal contrasts, are worried about multiplicity, and a replication is not feasible, then Bonferroni would be fine.

On the other hand, if one is content leaving the error rate at the per comparison level of $\alpha = 0.05$, then the usual contrasts can be computed and tested against the usual critical values of the *t* distribution. Of course, the multiplicity problem would be present and the researcher should replicate the results as a check against Type I errors.

I constructed a flow chart to guide your decision making.

(h) Some nice philosophical issues

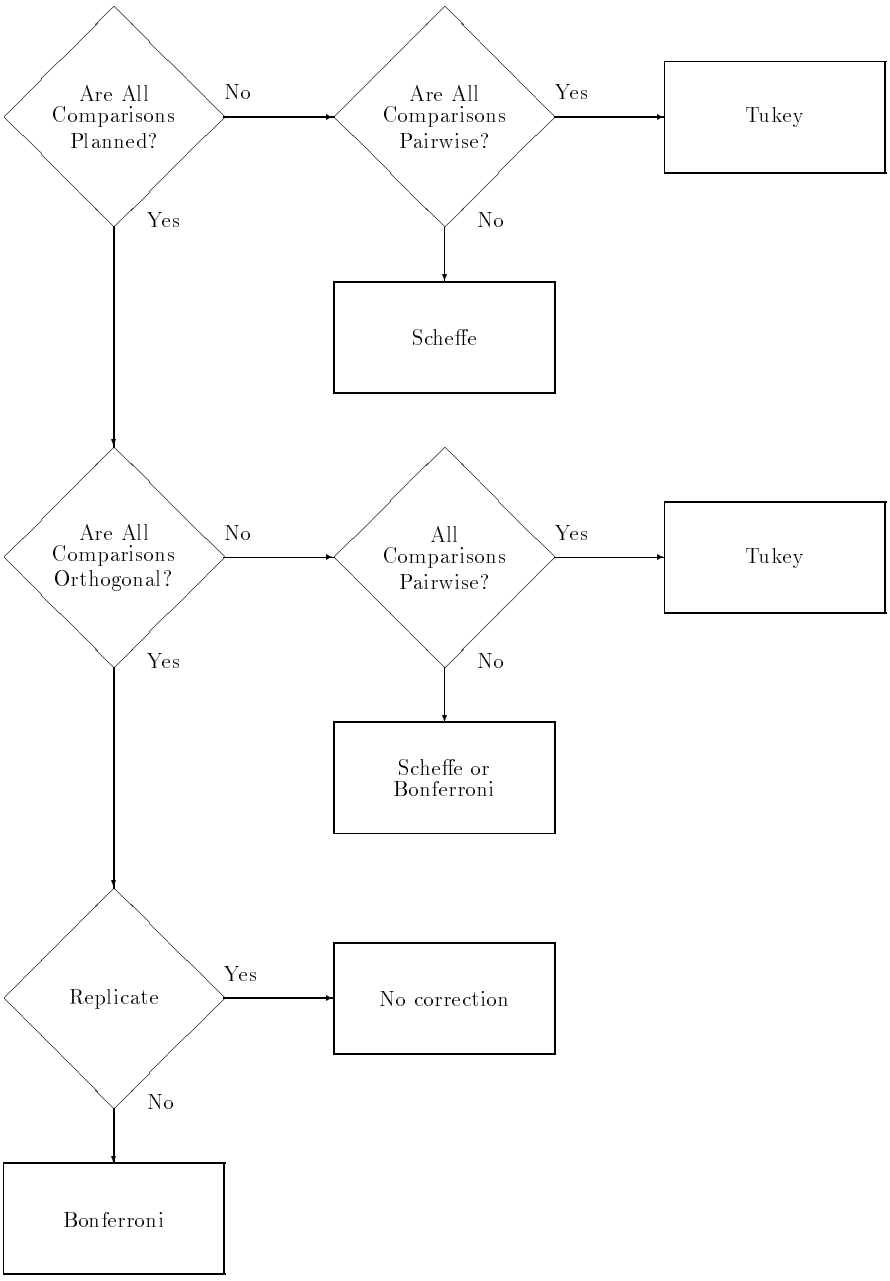
Let's pause for a moment and consider some philosophical issues. What does it mean not to have planned a set of comparisons? Was the experimenter in some kind of existential haze when designing and planning the study that prevented him from having hypotheses? Did the experimenter simply slap some groups together without much thought?

Also, how are we to interpret "planned"? Do we interpret it in the strict sense, where the direction of the means are stated prior to the data (one-tailed test); or in a weaker sense, where one "plans" to test whether the difference is different than zero (two-tailed) and tests whether the prediction is in the correct or incorrect direction?

There is something weird about one-tailed tests. Consider two researchers with competing theories. Each sets out to show her theory correct. Both of them independently conduct the same study. One plans to test the research hypothesis

$$\mathrm{H_a}: \quad \mu_1 > \mu_2,$$

FLOW CHART FOR COMPARISONS–Rich Gonzalez



the other plans to test the research hypothesis

$$H_a: \quad \mu_1 < \mu_2.$$

Both of them observe in their data that $\overline{Y}_1 > \overline{Y}_2$. The first researcher predicts this ordering so following the logic of a one-tailed test she can do her contrast and report a significant result. The second researcher predicted the reverse ordering so she, strictly following the classical approach to one-tailed tests, can only say she fails to reject the null hypothesis and must ignore the result in the opposite direction. Should the second researcher turn her back on her observation just because she didn't predict it? These kinds of problems have led to wide-spread debate about the merits of one- and two-tailed tests. It appears that in science we should always adopt a two-tailed criterion because we want to consider results that go against our predictions. There is much confusion about one vs. two tailed tests.

However, most people agree that in the extreme case of "fishing" corrective measures must be applied (e.g., "I'll try comparing all possible pairwise contrasts to see what's significant and report anything that is significantly different"). With so many different options, life becomes confusing. Between replication, Tukey, Bonferroni, and Scheffe, you now have the major tests in your toolbox. I'll give some heuristics to choose these various methods in different settings.

(i) Examples using the sleep deprivation data

Here is the SPSS syntax for performing post hoc pairwise tests. The ONEWAY command does it quite easily:

```
ONEWAY dv BY  group
 /RANGES TUKEY
 /RANGES SNK
 /RANGES SCHEFFE.
```

Some versions of SPSS may require separate ONEWAY calls for each /RANGE line. Different versions of SPSS format the output differently. First I'll present the format given by the most recent version of SPSS on the PC. Then I'll give alternate versions that you might find on other platforms.

Multiple Comparisons
Dependent Variable: DV

	(I) GROUP	(J) GROUP	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Tukey HSD	12hr	24hr	-1.3750	.6178	.141	-3.0618	.3118
		36hr	-3.2500(*)	.6178	.000	-4.9368	-1.5632
		48hr	-6.8750(*)	.6178	.000	-8.5618	-5.1882
	24hr	12hr	1.3750	.6178	.141	-.3118	3.0618
		36hr	-1.8750(*)	.6178	.025	-3.5618	-.1882
		48hr	-5.5000(*)	.6178	.000	-7.1868	-3.8132
	36hr	12hr	3.2500(*)	.6178	.000	1.5632	4.9368
		24hr	1.8750(*)	.6178	.025	.1882	3.5618
		48hr	-3.6250(*)	.6178	.000	-5.3118	-1.9382
	48hr	12hr	6.8750(*)	.6178	.000	5.1882	8.5618
		24hr	5.5000(*)	.6178	.000	3.8132	7.1868
		36hr	3.6250(*)	.6178	.000	1.9382	5.3118

DV

	GROUP	N	Subset for alpha = .050			
			1	2	3	4
Tukey HSD(a)	12hr	8	19.3750			
	24hr	8	20.7500			
	36hr	8		22.6250		
	48hr	8			26.2500	
	Sig.		.141	1.000	1.000	
Student-Newman-Keuls(a)	12hr	8	19.3750			
	24hr	8		20.7500		
	36hr	8			22.6250	
	48hr	8				26.2500
	Sig.		1.000	1.000	1.000	1.000
Scheffe(a)	12hr	8	19.3750			
	24hr	8	20.7500			
	36hr	8		22.6250		
	48hr	8			26.2500	
	Sig.		.200	1.000	1.000	

Means for groups in homogeneous subsets are displayed.

a Uses Harmonic Mean Sample Size = 8.000.

```
EXAMPLE OF TUKEY'S W USING SLEEP DEPRIVATION DATA (ALTERNATE OUTPUT)

 TUKEY-HSD PROCEDURE
 RANGES FOR THE 0.050 LEVEL -

          3.86   3.86   3.86

 THE RANGES ABOVE ARE TABLE RANGES.
 THE VALUE ACTUALLY COMPARED WITH MEAN(J)-MEAN(I) IS..
        0.8737 * RANGE * DSQRT(1/N(I) + 1/N(J))

   (*) DENOTES PAIRS OF GROUPS SIGNIFICANTLY DIFFERENT AT THE 0.050 LEVEL

                         G G G G
                         r r r r
                         p p p p

     Mean      Group      1 2 3 4

    19.3750    Grp 1
    20.7500    Grp 2
    22.6250    Grp 3      * *
    26.2500    Grp 4      * * *


  HOMOGENEOUS SUBSETS   (SUBSETS OF GROUPS, WHOSE HIGHEST AND LOWEST MEANS
                         DO NOT DIFFER BY MORE THAN THE SHORTEST
                         SIGNIFICANT RANGE FOR A SUBSET OF THAT SIZE)

SUBSET  1
GROUP       Grp 1         Grp 2
MEAN       19.3750       20.7500
- - - - - - - - - - - - - - - - - -

SUBSET  2
GROUP       Grp 3
MEAN       22.6250
- - - - - - - - - -

SUBSET  3
GROUP       Grp 4
MEAN       26.2500
- - - - - - - - - -
```

```
EXAMPLE SHOWING SNK TEST USING THE SLEEP DEPRIVATION DATA (ALTERNATE OUTPUT)

 STUDENT-NEWMAN-KEULS PROCEDURE
 RANGES FOR THE 0.050 LEVEL -
          2.90   3.49   3.86

 THE RANGES ABOVE ARE TABLE RANGES.
 THE VALUE ACTUALLY COMPARED WITH MEAN(J)-MEAN(I) IS..
        0.8737 * RANGE * DSQRT(1/N(I) + 1/N(J))

   (*) DENOTES PAIRS OF GROUPS SIGNIFICANTLY DIFFERENT AT THE 0.050 LEVEL

                         G G G G
                         r r r r
                         p p p p

     Mean      Group      1 2 3 4

    19.3750    Grp 1
    20.7500    Grp 2      *
    22.6250    Grp 3      * *
    26.2500    Grp 4      * * *

  HOMOGENEOUS SUBSETS   (SUBSETS OF GROUPS, WHOSE HIGHEST AND LOWEST MEANS
                         DO NOT DIFFER BY MORE THAN THE SHORTEST
                         SIGNIFICANT RANGE FOR A SUBSET OF THAT SIZE)
SUBSET  1
GROUP       Grp 1
MEAN       19.3750
- - - - - - - - - -
SUBSET  2
GROUP       Grp 2
MEAN       20.7500
- - - - - - - - - -
SUBSET  3
GROUP       Grp 3
MEAN       22.6250
- - - - - - - - - -
SUBSET  4
GROUP       Grp 4
MEAN       26.2500
```

```
EXAMPLE OF THE SCHEFFE TEST (PAIRWISE) USING THE SLEEP DEPRIVATION DATA
(ALTERNATE OUTPUT)

 SCHEFFE PROCEDURE
 RANGES FOR THE 0.050 LEVEL -

          4.20   4.20   4.20

 THE RANGES ABOVE ARE TABLE RANGES.
 THE VALUE ACTUALLY COMPARED WITH MEAN(J)-MEAN(I) IS..
        0.8737 * RANGE * DSQRT(1/N(I) + 1/N(J))

   (*) DENOTES PAIRS OF GROUPS SIGNIFICANTLY DIFFERENT AT THE 0.050 LEVEL

                         G G G G
                         r r r r
                         p p p p

     Mean      Group      1 2 3 4

    19.3750    Grp 1
    20.7500    Grp 2
    22.6250    Grp 3      * *
    26.2500    Grp 4      * * *


  HOMOGENEOUS SUBSETS   (SUBSETS OF GROUPS, WHOSE HIGHEST AND LOWEST MEANS
                         DO NOT DIFFER BY MORE THAN THE SHORTEST
                         SIGNIFICANT RANGE FOR A SUBSET OF THAT SIZE)

 SUBSET  1
 GROUP       Grp 1         Grp 2
 MEAN       19.3750       20.7500
 - - - - - - - - - - - - - - - - - -
 SUBSET  2
 GROUP       Grp 3
 MEAN       22.6250
 - - - - - - - - - -
 SUBSET  3
 GROUP       Grp 4
 MEAN       26.2500
 - - - - - - - - - -
```

Near the top of each section of output appears three numbers (e.g., at the top of the Tukey output we see 3.86, 3.86, 3.86). These numbers refer to the Tukey tabled values. Note that the SNK test has the numbers 2.90, 3.49, and 3.86 because the criterion (and hence the table value) changes as a function of the number of steps. The Scheffe S is 4.20 regardless of steps. In this example, Tukey is less conservative than Scheffe (i.e., the number to beat is not as large).

An explanation of SPSS. It is instructive to compare the formula printed in the SPSS output with the formula given in Equation 3-21 for unequal sample sizes. SPSS writes the formula

$$.8737 * q * \sqrt{1/\mathrm{n}_i + 1/\mathrm{n}_j} \tag{3-26}$$

where q is a "table lookup" from the studentized range statistics table. The number

.8737 is specific to these data. It is $\sqrt{\frac{\text{MSE}}{2}} = \sqrt{\frac{1.5268}{2}}$. You can see that the formula given in the SPSS output is identical to W given in the lecture notes. To find W for this example, just plug in the numbers

$$W = .8737 * 3.86 * \sqrt{1/8 + 1/8} \quad (3\text{-}27)$$
$$= 1.686 \quad (3\text{-}28)$$

So, any observed mean difference that exceeds W is statistically significant by Tukey.

For the SNK test, SPSS uses the same formula as that for Tukey's test but adjusts the value taken from the studentized range table (i.e., the value printed "range" in the output). For the SNK, the value with the correct number of steps is used. For the Scheffe test, which is based on the F distribution rather than the studentized range distribution, SPSS uses the F distribution.

To check your understanding, I suggest you perform TUKEY, SNK and SCHEFFE by hand on the sleep deprivation data and compare your "hand results" with the SPSS results presented here.

(j) Another example showing the use of contrast and post hoc tests

Imagine that a psychologist has performed a study to compare three different treatments for alleviating agoraphobia. Twenty-four subjects have been randomly assigned to one of four conditions: control group, a psychodynamic treatment, behavioral treatment A, and behavioral treatment B. The following posttest scores were obtained on a fear scale, where lower scores indicate worse phobia.

The raw data are:

Control	Psycho	Beh A	Beh B
5	3	6	6
3	7	5	9
1	6	7	9
4	3	5	4
3	4	3	5
5	7	4	6

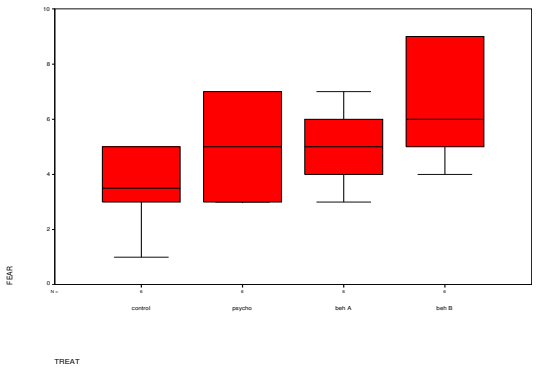
First, we examine assumptions. SPSS commands follow.

data list file = 'data.clinic' free / treat fear

value labels treat 1 'control' 2 'psycho' 3 'beh A' 4 'beh B'

examine variables fear by treat
/plot boxplot npplot .

I'll only show the boxplots in the interest of space. You can generate the normal plots on your own.



The equality of variance assumption appears satisfied, though difficult to tell with only six subjects per cell. No extreme outliers are evident. Symmetry seems okay too.

Next, examine the structural model and parameter estimates.

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij} \tag{3-29}$$

The parameter estimates are easy to calculate. All you need are the cell means and the grand mean. Be careful that the grand mean is the "real" grand mean (the mean of the cell means rather than the mean of all the data). This information appears in several places in the SPSS output (i.e., you'll get it when you ask for boxplots, and if you use the **/statistics = descriptives** subcommand in **ONEWAY**, you can ask for **MEANS = tables fear by treat.**) Below I include the output from the **MEANS** command.

means tables = fear by treat .

Variable	Value	Label	Mean	Std Dev	Cases	Parameter est.
For Entire Population			5.0000	1.9560	24	$\hat{\mu} = 5$
TREAT	1.00	control	3.5000	1.5166	6	$\hat{\alpha}_1 = \overline{Y}_1 - \hat{\mu} = -1.5$
TREAT	2.00	psycho	5.0000	1.8974	6	$\hat{\alpha}_2 = \overline{Y}_2 - \hat{\mu} = 0$
TREAT	3.00	beh A	5.0000	1.4142	6	$\hat{\alpha}_3 = \overline{Y}_3 - \hat{\mu} = 0$
TREAT	4.00	beh B	6.5000	2.0736	6	$\hat{\alpha}_4 = \overline{Y}_4 - \hat{\mu} = 1.5$

We can see that the psychodynamic and beh A groups had treatment effects of zero; the two remaining groups had treatment effects of 1.5 (note sign).

Next we run the inferential tests: ANOVA & contrasts

I'll use the ONEWAY command for the overall ANOVA, contrasts, and post hoc tests. Note that with only six subjects per cell we shouldn't be too optimistic that this study will lead to significant results. Usually, effect sizes are small enough that six subjects per cell doesn't give one much power.

```
oneway fear by treat
 /statistics all
 /contrasts = -3, 1, 1, 1
 /contrasts =  0, -2, 1, 1
 /contrasts =  0, 0, 1, -1
 /ranges=tukey .
```

ANALYSIS OF VARIANCE

SOURCE	D.F.	SUM OF SQUARES	MEAN SQUARES	F RATIO	F PROB.
BETWEEN GROUPS	3	27.0000	9.0000	2.9508	.0575
WITHIN GROUPS	20	61.0000	3.0500		
TOTAL	23	88.0000			

This ANOVA tells us that the omnibus test is not statistically significant at $\alpha = 0.05$. That is, the four means are not different from each other. But this test is not very informative as to where the differences could be (if there were any) because it is examining the four means as a set.

We need to be careful because there are only six subjects per cell so this is not a powerful

test.

The design calls for a natural set of contrasts. Make sure you understand why these three contrasts provide a natural set of comparisons for this particular study. These three contrasts are orthogonal.

```
 CONTRAST COEFFICIENT MATRIX

            Grp 1       Grp 3
                  Grp 2       Grp 4

CONTRAST  1  -3.0   1.0   1.0   1.0
CONTRAST  2   0.0  -2.0   1.0   1.0
CONTRAST  3   0.0   0.0   1.0  -1.0


                                        POOLED VARIANCE ESTIMATE
                    VALUE      S. ERROR       T VALUE    D.F.       T PROB.
CONTRAST  1         6.0000       2.4698         2.429    20.0         0.025
CONTRAST  2         1.5000       1.7464         0.859    20.0         0.401
CONTRAST  3        -1.5000       1.0083        -1.488    20.0         0.152

                                      SEPARATE VARIANCE ESTIMATE
                    VALUE      S. ERROR       T VALUE    D.F.       T PROB.
CONTRAST  1         6.0000       2.2583         2.657    10.1         0.024
CONTRAST  2         1.5000       1.8574         0.808     9.3         0.439
CONTRAST  3        -1.5000       1.0247        -1.464     8.8         0.178
```

Next, I perform a 95% CI on Contrast #1 to illustrate the computation. From the output we see that $\hat{I} = 6$ and se($\hat{I}$) = 2.47. The *t*-value for a 95% CI with 20 degrees of freedom is 2.086 (from the *t*-table). Recall that the formula for a CI on a contrast is

$$\begin{aligned} \hat{I} &\pm t_{\alpha/2,\text{df}}\ \text{se}(\hat{I}) \\ 6 &\pm (2.086)(2.47) \\ 6 &\pm 5.15 \end{aligned}$$

yielding the interval (0.85, 11.15). This interval does not contain zero.

Bayesian and Bootstrap Contrasts

To run Bayesian tests of contrasts in SPSS we need to first learn how to recode ANOVA into regression because currently regression is the only way to run contrasts in SPSS using the Bayesian approach. Basically, each orthogonal contrast becomes a predictor and then each coefficient in that regression estimates "i hat." We'll come back to this later in the semester after covering how to implement ANOVA in a regression context. Bayesian contrast testing in R is easier and I show that in the appendix in these lecture notes.

Bootstrapping contrasts in SPSS is easy. Just type the bootstrap code before the usual ONEWAY command as in

```
BOOTSTRAP
  /SAMPLING METHOD=SIMPLE
  /VARIABLES TARGET=dv INPUT=group
  /CRITERIA CILEVEL=95 CITYPE=PERCENTILE  NSAMPLES=1000
  /MISSING USERMISSING=EXCLUDE.
ONEWAY dv BY group
 /contrast 3 -1 -1 -1
 /contrast 0 2 -1 -1
 /contrast 0 0 1 -1.
```

This yields the same table of contrast tests one usually sees for ONEWAY but with an extra column giving the p values computed through the bootstrap procedure.

In R the bootstrap is easy and I show it in the appendix.

Multiplicity of tests

One could merely report these three contrast values and corresponding t values. That would be fine. But the problem of multiplicity is present and the researcher would want to replicate the results. If a replication is not feasible, then one solution to the multiplicity problem is to use a Bonferroni-type correction to control the experimentwise error rate. Because there are three contrasts, the per comparison α' will be $\frac{\alpha}{3} = 0.017$. None of the three contrasts are now significant because 0.017 is the tail probability to beat, not 0.05. Note that the Bonferroni correction involves only a change in the criterion for what is deemed "statistically significant"—no other computational changes are necessary when implementing the Bonferroni correction.

Post hoc tests

Let's imagine that the researcher didn't have any hypotheses whatsoever. A strange statement given the specifics of the study. Because the researcher is on a fishing expedition, then *post hoc* tests would be appropriate. Here I show all possible pair-wise comparisons using Tukey's Honestly Significant Difference procedure.

```
POST HOC TEST (ALL POSSIBLE PAIRWISE COMPARISONS)
TUKEY-HSD PROCEDURE
RANGES FOR THE 0.050 LEVEL -

          3.95   3.95   3.95
```

```
THE RANGES ABOVE ARE TABLE RANGES.
THE VALUE ACTUALLY COMPARED WITH MEAN(J)-MEAN(I) IS..
        1.2349 * RANGE * DSQRT(1/N(I) + 1/N(J))

   (*) DENOTES PAIRS OF GROUPS SIGNIFICANTLY DIFFERENT AT THE 0.050 LEVEL


                        G G G G
                        r r r r
                        p p p p

     Mean      Group     1 2 3 4

     3.5000    Grp 1
     5.0000    Grp 2
     5.0000    Grp 3
     6.5000    Grp 4     *


  HOMOGENEOUS SUBSETS   (SUBSETS OF GROUPS, WHOSE HIGHEST AND LOWEST MEANS
                         DO NOT DIFFER BY MORE THAN THE SHORTEST
                         SIGNIFICANT RANGE FOR A SUBSET OF THAT SIZE)
SUBSET  1

GROUP       Grp 1         Grp 2         Grp 3
MEAN        3.5000        5.0000        5.0000
- - - - - - - - - - - - - - - - - - - - - - - -
SUBSET  2

GROUP       Grp 2         Grp 3         Grp 4
MEAN        5.0000        5.0000        6.5000
- - - - - - - - - - - - - - - - - - - - - - - -
```

I now illustrate the computation of the Tukey test. Using the method for W with equal n we have

$$\begin{aligned} W &= q_{\alpha}(T,v)\sqrt{\frac{MSE}{n}} && (3\text{-}30)\\ &= 3.95\sqrt{\frac{3.05}{6}} && (3\text{-}31)\\ &= 2.82 && (3\text{-}32) \end{aligned}$$

So any pairwise difference between means that exceeds W=2.82 is statistically significant by Tukey. In this example, only the difference between Group 1 and Group 4 means exceeds 2.82.

To illustrate the Scheffe test I'll use the SPSS shortcut that I introduced earlier in these lecture notes. Recall that we setup a new t critical based on the Scheffe test, and we use that new t critical to evaluate the observed t in the contrast portion of the SPSS output. Recall the new t critical is given by

$$\begin{aligned} t_{\text{critical}} &= \sqrt{(T\text{-}1)F_{\alpha,\ df1,\ df2}}\\ &= \sqrt{(4-1)*3.098} \end{aligned}$$

$$= \quad 3.049$$

The tabled F for 3 and 20 degrees of freedom is 3.098. We scan the contrast output and look for contrasts that have an observed t that exceeds 3.049. None of the contrasts are significant by Scheffe. Note that the usual t criterion for an uncorrected t is a number near 2 (depending on degrees of freedom). The Scheffe is quite conservative at 3.049.

Another equivalent way to compute the Scheffe test is to calculate S, which is the contrast value to beat. In this example, SPSS computes $\hat{\text{I}}$ and its standard error. I'll just illustrate the first contrast in this example. Recall the formula

$$\begin{aligned} \text{S} &= \sqrt{\text{V}(\hat{\text{I}})}\sqrt{(\text{T-1})\text{F}_{\alpha,\ \text{df1},\ \text{df2}}} && \text{(3-33)} \\ &= 2.4698 * 3.049 && \text{(3-34)} \\ &= 7.53 && \text{(3-35)} \end{aligned}$$

where the 2.4698 is the se printed in the output and the 3.049 is the term I showed you how to compute in the previous paragraph. So any observed $\hat{\text{I}}$ that exceeds 7.53 is statistically significant by Scheffe. None of the observed $\hat{\text{I}}$s exceed 7.53 so, as we saw with the other equivalent methods, none of the three contrasts are statistically significant.

To compute the Welch-like version of the Scheffe just replace df2 (degrees of freedom for the denominator) with the degrees of freedom reported in the "separate variance" portion of the SPSS output and use the separate variance estimate of the standard error of $\hat{\text{I}}$ (or for the other method, the separate variance observed t). You use this newer degrees of freedom in the table lookup of F.

Why not perform the contrast (1, 0, 0, -1) corresponding to the comparison between control and Behavioral Treatment B? It depends on how you decided to test this contrast. Did you predict that the only difference would be between the control and Beh B conditions? If so, the you can go ahead and run such a contrast. The contrast would be tested at $\alpha = 0.5$. What are the remaining two orthogonal contrasts? If you run those tests, you will need to grapple with the multiplicity problem and need to decide whether or not to perform a Bonferroni correction.

However, if you look at the data and say, "Gee, that's interesting. I would have never guessed that Beh B would have been the only group that would be different than the control." Then you should perform a *post hoc* test such as Scheffe to take into account chance factors that might have lead to such a result.

(k) Uniqueness of a contrast

Contrasts are unique up to a scale transformation. You get the identical *p*-value if you multiply all the contrast weights by a constant real number (positive or negative). The

size of the CI will shrink or expand accordingly, but key features (such as whether or not the CI includes 0 and whether or not the CI overlaps with other contrasts that are on the same scale) remain the same.

Let's look at Contrast #1 in the previous example. What does the value $\hat{I} = 6$ mean? It means that 3 times the control group minus the sum of the three treatment means equals 6. Some people prefer using contrasts that are normalized so the sum of the absolute values of the weights equals 2 (i.e., $\sum |a| = 2$).

For Contrast #1 above I used the weights (-3, 1, 1, 1). I could have used (-1, $\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$) and would have seen an identical *p*-value. The weights in both contrasts are proportional so they test exactly the same hypothesis. So, I know, without having to use any SPSS commands, that the value of the contrast (-1, $\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$) will be 2 (i.e., $\frac{6}{3}$) and the standard error will be 0.8233 (i.e,. $\frac{2.4698}{3}$).

To check this (and to convince you by example) here is the SPSS command for this new contrast. Unfortunately, SPSS does not allow one to enter fractions as contrast weights so one needs to round off. Let's see what happens when I enter the contrast (-1, 0.33, 0.33, 0.34).

```
data list file = 'data.clinic' free /  treat fear

value labels treat 1 'control' 2 'psycho' 3 'beh A' 4 'beh B'

oneway fear by treat
 /statistics all
 /contrasts =  -1, 0.33, 0.33, 0.34 .
```

```
                Grp 1       Grp 3
                      Grp 2       Grp 4
CONTRAST  1  -1.0   0.3   0.3   0.3
                                          POOLED VARIANCE ESTIMATE
                   VALUE      S. ERROR       T VALUE    D.F.       T PROB.
CONTRAST  1*      2.0100        0.8233         2.441    20.0         0.024

                                          SEPARATE VARIANCE ESTIMATE
                              S. ERROR       T VALUE    D.F.       T PROB.
                                0.7535         2.667    10.1         0.023

          * ABOVE INDICATES SUM OF COEFFICIENTS IS NOT ZERO.
```

As we knew already, the contrast value is 2 (roundoff error) and the standard error is 0.8233. The 95% CI is (the *t*-value = 2.086 remains the same as before)

$$\begin{array}{lll} \hat{I} & \pm & t_{\alpha/2,\text{df}}\ \text{se}(\hat{I}) \\ 2 & \pm & (2.086)(0.8233) \\ 2 & \pm & 1.7174 \end{array}$$

yielding the interval (0.28, 3.72). Note that, as before, zero is not included.

Let's look at another example showing that contrasts are unique up to a scale transformation. Using the sleep deprivation data I presented in a previous lecture. All three contrasts are identical and yield identical p-values; the weights are proportional.

```
data list file = name free /  id dv codes rdv

value labels codes 1 '12hr' 2 '24hr' 3 '36hr' 4 '48hr'

oneway dv by codes
 /contrasts =  1, 1, -2, 0
 /contrasts =  -1, -1, 2, 0
 /contrasts =  .5, .5, -1, 0 .
```

ANALYSIS OF VARIANCE

SOURCE	D.F.	SUM OF SQUARES	MEAN SQUARES	F RATIO	F PROB.
BETWEEN GROUPS	3	213.2500	71.0833	46.5575	.0000
WITHIN GROUPS	28	42.7500	1.5268		
TOTAL	31	256.0000			

		Grp 1	Grp 2	Grp 3	Grp 4
CONTRAST	1	1.0	1.0	-2.0	0.0
CONTRAST	2	-1.0	-1.0	2.0	0.0
CONTRAST	3	0.5	0.5	-1.0	0.0

POOLED VARIANCE ESTIMATE

		VALUE	S. ERROR	T VALUE	D.F.	T PROB.
CONTRAST	1	-5.1250	1.0701	-4.789	28.0	0.000
CONTRAST	2	5.1250	1.0701	4.789	28.0	0.000
CONTRAST	3	-2.5625	0.5350	-4.789	28.0	0.000

SEPARATE VARIANCE ESTIMATE

S. ERROR	T VALUE	D.F.	T PROB.
1.0426	-4.916	14.5	0.000
1.0426	4.916	14.5	0.000
0.5213	-4.916	14.5	0.000

9. What to write in a results section?

Contrasts are easy to write because they should directly correspond to your hypotheses. So if you have three hypotheses, you can state each one to remind the reader, refer them to a table or figure of means and error bars, and then provide the results of the contrast analysis including t, df, se, pvalue and confidence interval for the contrast value $\hat{\text{I}}$. Example (sure, the writing could be improved but it serves to illustrate the key pieces of information that should be reported):

> .... Our second hypothesis is that the two experimental groups will lead to faster performance than the control condition. The three group means with bars depicting one standard error are presented in Figure XXX. The hypothesis was tested with the (1, 1, -2) contrast, which was statistically significant, t(215) = 2.45, p = .015, 95% CI for contrast value was (1.2, 4.5). Finally, the third hypothesis is that the group receiving experimental drug A will lead to faster performance than the group receiving experimental drug B. This hypothesis was tested with the (1, -1, 0) contrast, which was not statistically significant, t(df) = XX, p = YY, 95% CI for contrast value (a, z).

One could also add an effect size measure for the contrast to the long list of key pieces of information. One standard effect size measure is the R^2 for the contrast (basically the ratio of the sum of squares for that particular contrast over sum of squares total).

Post hoc tests usually involve some thinking about how best to present because the complexity of the design and the number of significant findings can lead to challenges for communicating the results succinctly. After reporting which post-hoc tests you used (Tukey, Scheffe, etc) you simply describe the pattern or refer to a figure that includes the post hoc results. Examples (assuming a table or figure of means with error bars are already presented).

> Post hoc Tukey tests revealed that none of the pairwise differences reached statistically significant.

> Post hoc Tukey tests revealed only two significant pairwise comparisons: Group D vs Group F, and Group B vs Group F.

No need to report specific test statistic values for these post hoc tests as the point is to report which pairwise differences emerge after controlling the Type I error rate. The means and their error bars are sufficient.

> The results of post hoc Tukey tests are depicted in Figure YYY. Horizontal line segments with an asterisk are statistically significant by the Tukey test. (Figure TBA in LN)

Appendix 1

```
ANOVA ON SLEEP DEPRIVATION DATA

SPSS SYNTAX:

oneway dv by codes
 /contrast 3 -1 -1 -1
 /contrast 0  2 -1 -1
 /contrast 0  0  1 -1
 /contrast 1  0  0 -1.
```

SOURCE	D.F.	SUM OF SQUARES	MEAN SQUARES	F RATIO	F PROB.
BETWEEN GROUPS	3	213.2500	71.0833	46.5575	.0000
WITHIN GROUPS	28	42.7500	1.5268		
TOTAL	31	256.0000			

CONTRAST COEFFICIENT MATRIX

		Grp 1	Grp 2	Grp 3	Grp 4
CONTRAST	1	3.0	-1.0	-1.0	-1.0
CONTRAST	2	0.0	2.0	-1.0	-1.0
CONTRAST	3	0.0	0.0	1.0	-1.0
CONTRAST	4	1.0	0.0	0.0	-1.0

POOLED VARIANCE ESTIMATE

		VALUE	S. ERROR	T VALUE	D.F.	T PROB.
CONTRAST	1	-11.5000	1.5133	-7.599	28.0	0.000
CONTRAST	2	-7.3750	1.0701	-6.892	28.0	0.000
CONTRAST	3	-3.6250	0.6178	-5.867	28.0	0.000
CONTRAST	4	-6.8750	0.6178	-11.128	28.0	0.000

SEPARATE VARIANCE ESTIMATE

		VALUE	S. ERROR	T VALUE	D.F.	T PROB.
CONTRAST	1	-11.5000	1.4745	-7.799	12.6	0.000
CONTRAST	2	-7.3750	1.0969	-6.724	13.5	0.000
CONTRAST	3	-3.6250	0.6178	-5.867	13.9	0.000
CONTRAST	4	-6.8750	0.6178	-11.128	13.9	0.000

The pooled variance estimate t values correspond to the formulae in the lecture. The separate variance estimate is a generalization of Welch's work to contrasts. Note that the separate variance estimate "corrects" for violations of the equality of variance assumption by reducing the degrees of freedom.

SSC for contrast 1 is given by (this example has equal sample sizes so the n_i are equal)

$$\begin{aligned} \text{SSC}_1 &= \frac{\hat{l}^2}{\sum \frac{a_i^2}{n_i}} \\ &= \frac{[(3)\overline{Y}_1 + (-1)\overline{Y}_2 + (-1)\overline{Y}_3 + (-1)\overline{Y}_4]^2}{\sum \frac{(3)^2+(-1)^2+(-1)^2+(-1)^2}{n}} \\ &= \frac{[(3)19.375 + (-1)20.75 + (-1)22.625 + (-1)26.25]^2}{\frac{12}{8}} \\ &= 88.167 \end{aligned}$$

Similarly, we can compute

$$\text{SSC}_2 = 72.52$$

$$\begin{aligned} \text{SSC}_3 &= 52.56 \\ \text{SSC}_4 &= 188.79 \end{aligned}$$

Note that SSB $=$ $\text{SSC}_1 + \text{SSC}_2 + \text{SSC}_3$. So what about SSC_4? It is not orthogonal with the other three contrasts and is consequently redundant. Note that SSC_4 accounts for quite a hefty chunk of the SSB.

Here is a source table reflecting the decomposition of SSB into the separate SSC.

```
SHOW BREAKDOWN OF SSB WITHIN THE ANOVA TABLE

                                 SUM OF          MEAN            F      t
        SOURCE           D.F.    SQUARES        SQUARES         RATIO value

 BETWEEN GROUPS             3      213.2500       71.0833      46.5575
      CONTRAST  1             1         88.17       88.17       57.75  7.60
      CONTRAST  2             1         72.52       72.52       47.49  6.89
      CONTRAST  3             1         52.56       52.56       34.42  5.87
 WITHIN GROUPS             28       42.7500        1.5268
 TOTAL                     31      256.0000
```

A SECOND EXAMPLE USING A DIFFERENT SET OF ORTHOGONAL CONTRASTS.
SLEEP DEPRIVATION DATA .

CONTRAST COEFFICIENT MATRIX (NOTE THAT ALL THREE ARE ORTHOGONAL)

	Grp 1	Grp 2	Grp 3	Grp 4
CONTRAST 1	-1.0	-1.0	-1.0	3.0
CONTRAST 2	-1.0	-1.0	2.0	0.0
CONTRAST 3	1.0	-1.0	0.0	0.0

POOLED VARIANCE ESTIMATE

	VALUE	S. ERROR	T VALUE	D.F.	T PROB.
CONTRAST 1	16.0000	1.5133	10.573	28.0	0.000
CONTRAST 2	5.1250	1.0701	4.789	28.0	0.000
CONTRAST 3	-1.3750	0.6178	-2.226	28.0	0.034

SEPARATE VARIANCE ESTIMATE

	VALUE	S. ERROR	T VALUE	D.F.	T PROB.
CONTRAST 1	16.0000	1.5512	10.315	11.5	0.000
CONTRAST 2	5.1250	1.0426	4.916	14.5	0.000
CONTRAST 3	-1.3750	0.6178	-2.226	13.9	0.043

SHOW BREAKDOWN OF SSB

SOURCE	D.F.	SUM OF SQUARES	MEAN SQUARES	F RATIO	t value
BETWEEN GROUPS	3	213.2500	71.0833	46.5575	
CONTRAST 1	1	170.67	170.67	111.782	10.573
CONTRAST 2	1	35.02	35.02	22.94	4.79
CONTRAST 3	1	7.56	7.56	4.953	2.226
WITHIN GROUPS	28	42.7500	1.5268		
TOTAL	31	256.0000			

A THIRD EXAMPLE USING A DIFFERENT SET OF ORTHOGONAL CONTRASTS (POLYNOMIAL).

CONTRAST COEFFICIENT MATRIX (NOTE THAT ALL THREE ARE ORTHOGONAL)

	Grp 1	Grp 2	Grp 3	Grp 4
CONTRAST 1	-3.0	-1.0	1.0	3.0
CONTRAST 2	1.0	-1.0	-1.0	1.0
CONTRAST 3	-1.0	3.0	-3.0	1.0

		POOLED VARIANCE ESTIMATE				
		VALUE	S. ERROR	T VALUE	D.F.	T PROB.
CONTRAST	1	22.5000	1.9537	11.517	28.0	0.000
CONTRAST	2	2.2500	0.8737	2.575	28.0	0.016
CONTRAST	3	1.2500	1.9537	0.640	28.0	0.527

		SEPARATE VARIANCE ESTIMATE				
		VALUE	S. ERROR	T VALUE	D.F.	T PROB.
CONTRAST	1	22.5000	1.9537	11.517	17.0	0.000
CONTRAST	2	2.2500	0.8737	2.575	27.8	0.016
CONTRAST	3	1.2500	1.9537	0.640	17.0	0.531

SHOW BREAKDOWN OF SSB

SOURCE	D.F.	SUM OF SQUARES	MEAN SQUARES	F RATIO	t value
BETWEEN GROUPS	3	213.2500	71.0833	46.5575	
CONTRAST 1	1	202.50	202.50	132.63	11.52
CONTRAST 2	1	10.13	10.13	6.63	2.57
CONTRAST 3	1	0.62	0.62	0.41	0.64
WITHIN GROUPS	28	42.7500	1.5268		
TOTAL	31	256.0000			

Appendix 2

ALL PAIRWISE COMPARISONS OF THE FOUR GROUPS IN THE SLEEP DEPRIVATION STUDY

CONTRAST COEFFICIENT MATRIX

		Grp 1	Grp 2	Grp 3	Grp 4
CONTRAST	1	1.0	-1.0	0.0	0.0
CONTRAST	2	1.0	0.0	-1.0	0.0
CONTRAST	3	1.0	0.0	0.0	-1.0
CONTRAST	4	0.0	1.0	-1.0	0.0
CONTRAST	5	0.0	1.0	0.0	-1.0
CONTRAST	6	0.0	0.0	1.0	-1.0

POOLED VARIANCE ESTIMATE

		VALUE	S. ERROR	T VALUE	D.F.	T PROB.
CONTRAST	1	-1.3750	0.6178	-2.226	28.0	0.034
CONTRAST	2	-3.2500	0.6178	-5.260	28.0	0.000
CONTRAST	3	-6.8750	0.6178	-11.128	28.0	0.000
CONTRAST	4	-1.8750	0.6178	-3.035	28.0	0.005
CONTRAST	5	-5.5000	0.6178	-8.902	28.0	0.000
CONTRAST	6	-3.6250	0.6178	-5.867	28.0	0.000

SEPARATE VARIANCE ESTIMATE

		VALUE	S. ERROR	T VALUE	D.F.	T PROB.
CONTRAST	1	-1.3750	0.6178	-2.226	13.9	0.043
CONTRAST	2	-3.2500	0.5939	-5.473	14.0	0.000
CONTRAST	3	-6.8750	0.6178	-11.128	13.9	0.000
CONTRAST	4	-1.8750	0.6178	-3.035	13.9	0.009
CONTRAST	5	-5.5000	0.6409	-8.582	14.0	0.000
CONTRAST	6	-3.6250	0.6178	-5.867	13.9	0.000

Recall that the sums of squares for each contrasts is given by

$$SSC_i = \frac{\hat{l}^2}{\sum \frac{a_i^2}{n_i}} \tag{3-36}$$

Note that for all these contrasts the denominator of Equation 3-36 is 0.25. The numerator is simply the "value" of the contrast squared.

The sums of squares for each contrast pairwise contrast are

$$\begin{aligned} SSC_1 &= 7.6 \\ SSC_2 &= 42.2 \\ SSC_3 &= 188.8 \\ SSC_4 &= 14.1 \\ SSC_5 &= 121.0 \\ SSC_6 &= 52.6 \end{aligned}$$

Appendix 3: R syntax

Given that there is so much special "R knowledge" to impart I decided to combine all the R syntax for this set of lecture notes in one appendix rather than sprinkle it throughout the notes like I do for SPSS. This way I can separate learning statistics from learning R.

There are many ways to test contrasts and post-hoc tests in R. I also work in some common R errors in this discussion, such as the case of incorrectly using the aov() command without a grouping code that is specifically defined as a factor.

Here I list a few of the basics. First, read data, label columns, create data.frame.

```
sleep.data <- read.table("/Users/gonzo/rich/Teach/Gradst~1/unixfiles/lectnotes/lect3/da
names(sleep.data) <- c("subject", "performance", "group",
    "residual")
sleep.data <- data.frame(sleep.data)
```

Run regular ANOVA, just prints the omnibus F test, no contrasts by default.

```
summary(aov(performance ~ group, data = sleep.data))
```

```
##             Df Sum Sq Mean Sq F value
## group        1  202.5  202.50   113.6
## Residuals   30   53.5    1.78
##               Pr(>F)
## group       1.02e-11 ***
## Residuals
## ---
## Signif. codes:
##   0 '***' 0.001 '**' 0.01 '*' 0.05
##   '.' 0.1 ' ' 1
```

Do you see something wrong with this source table? There are four levels of the group factor so there should be 3 degrees of freedom for between groups, but the output lists 1 degree of freedom. The problem is that R does not know that group is a factor; it interpreted the column labeled group as the numbers 1 to 4 rather than labels denoting four groups. We need to define factors explicitly; let's try that again.

```
sleep.data$group <- factor(sleep.data$group)
summary(aov(performance ~ group, data = sleep.data))
```

```
##             Df Sum Sq Mean Sq F value
## group        3 213.25   71.08   46.56
## Residuals   28  42.75    1.53
##               Pr(>F)
## group       5.22e-11 ***
## Residuals
## ---
```

```
## Signif. codes:
##   0 '***' 0.001 '**' 0.01 '*' 0.05
##   '.' 0.1 ' ' 1
```

These kinds of silly errors illustrate why it is important to know what you expect to see (such as the degrees of freedom) before running the command so you can spot errors. If you aren't careful you can end up publishing incorrect analyses, and unfortunately there are many such examples in the literature.

Contrasts in R

You can define contrasts directly in a factor. So if we have a grouping variable with four levels for the sleep deprivation example we first define group variable as a factor and then define an orthogonal set of contrasts for that factor. The benefit of this approach is that as you call that factor in various commands the "contrast structure" follows the variable and you don't have to keep re-specifying the contrasts. For example, if I want to test say the orthogonal set of three contrasts, c(1,-1,0,0), c(1,1,-2,0), and c(1,1,1,-3), the following two commands will do the trick.

```
contrasts(sleep.data$group) <- cbind(C12.24 = c(1,
    -1, 0, 0), C1224.36 = c(1, 1, -2, 0), C122436.48 = c(1,
    1, 1, -3))
# check the command worked as intended
contrasts(sleep.data$group)
```

```
##   C12.24 C1224.36 C122436.48
## 1      1        1          1
## 2     -1        1          1
## 3      0       -2          1
## 4      0        0         -3
```

```
anova.output <- aov(performance ~ group, data = sleep.data)
summary(anova.output, split = list(group = list(C12.24 = 1,
    C1224.36 = 2, C122436.48 = 3)))
```

```
##                     Df Sum Sq Mean Sq
## group                3 213.25   71.08
##   group: C12.24      1   7.56    7.56
##   group: C1224.36    1  35.02   35.02
##   group: C122436.48  1 170.67  170.67
## Residuals           28  42.75    1.53
##                     F value   Pr(>F)
## group                46.558 5.22e-11
##   group: C12.24       4.953   0.0343
##   group: C1224.36    22.938 4.93e-05
##   group: C122436.48 111.782 2.78e-11
## Residuals
##
## group               ***
##   group: C12.24     *
##   group: C1224.36   ***
```

```
##   group: C122436.48 ***
## Residuals
## ---
## Signif. codes:
##   0 '***' 0.001 '**' 0.01 '*' 0.05
##   '.' 0.1 ' ' 1
```

The syntax of the last line is interpreted as split the output of anova.output into the first, second and third contrasts that are defined in the factor(group) (i.e., the first contrast in the contrast(group) is assigned the label "C12.24", the second contrast in the contrast(group) is assigned the label "C1224.36," etc.). That's a lot of typing and I generally avoid the aov() command for reasons that I'll go into in LN4 due to the way the aov() treats factorial designs with unequal sample sizes. Note that this output is not in the form of the usual contrast value, se, t and p but rather in terms of the decomposition of the sum of squares (think pie chart). If you want to report the t rather than the F, just take the sqrt of the F.

```
sqrt(4.953)
```

```
## [1] 2.225534
```

```
sqrt(22.938)
```

```
## [1] 4.789363
```

```
sqrt(111.782)
```

```
## [1] 10.5727
```

We can do this because for F's with one degree of freedom in the numerator (all contrasts have one numerator df), we can use the relation that F^2 = t, the pvalues will be the same, etc.

Some R commands like lm() will use this contrast information in the factor and organize the output accordingly; this is the version I usually use because it is so simple and the output is in terms of the familiar contrast value, se, t and p. But for now you should only use it for orthogonal sets of contrasts as I show below.

```
output.lm <- lm(performance ~ group, sleep.data)
summary(output.lm)
```

```
##
## Call:
## lm(formula = performance ~ group, data = sleep.data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -2.3750 -0.6562  0.0000  0.6562  2.3750
##
## Coefficients:
##                 Estimate Std. Error
## (Intercept)      22.2500     0.2184
## groupC12.24      -0.6875     0.3089
## groupC1224.36    -0.8542     0.1783
## groupC122436.48  -1.3333     0.1261
##                 t value Pr(>|t|)
## (Intercept)     101.863  < 2e-16 ***
## groupC12.24      -2.226   0.0343 *
## groupC1224.36    -4.789 4.93e-05 ***
## groupC122436.48 -10.573 2.78e-11 ***
## ---
## Signif. codes:
##   0 '***' 0.001 '**' 0.01 '*' 0.05
##   '.' 0.1 ' ' 1
##
## Residual standard error: 1.236 on 28 degrees of freedom
## Multiple R-squared:  0.833,Adjusted R-squared:  0.8151
## F-statistic: 46.56 on 3 and 28 DF,  p-value: 5.222e-11
```

The t from the lm() command is identical to the sqrt of the F from the aov() command above.

You can get the source table from the output of the lm() command through the anova() command

```
anova(output.lm)


## Analysis of Variance Table
##
## Response: performance
##           Df Sum Sq Mean Sq F value
## group      3 213.25  71.083  46.557
## Residuals 28  42.75   1.527
##              Pr(>F)
## group     5.222e-11 ***
## Residuals
## ---
## Signif. codes:
##   0 '***' 0.001 '**' 0.01 '*' 0.05
##   '.' 0.1 ' ' 1
```

You can also test contrasts directly using the fit.contrast() command in the library(gmodels). This is probably the one you should use for now because it works for both orthogonal and nonorthogonal contrasts.

```
library(gmodels)
out.model <- aov(performance ~ group, sleep.data)
fit.contrast(out.model, "group", c(1, -1, 0, 0), df = T,
```

```
    conf.int = 0.95, show.all = T)


##                       Estimate
## group c=( 1 -1 0 0 )   -1.375
##                       Std. Error
## group c=( 1 -1 0 0 )  0.6178159
##                        t value
## group c=( 1 -1 0 0 ) -2.225582
##                          Pr(>|t|) DF
## group c=( 1 -1 0 0 ) 0.03427037 28
##                        lower CI
## group c=( 1 -1 0 0 ) -2.640538
##                          upper CI
## group c=( 1 -1 0 0 ) -0.1094616
## attr(,"class")
## [1] "fit_contrast"


# or if you want an entire orthogonal set
fit.contrast(out.model, "group", rbind(`12v24` = c(1,
    -1, 0, 0), `36v(12+24)` = c(1, 1, -2, 0), `48vall` = c(1,
    1, 1, -3)))


##                 Estimate Std. Error
## group12v24        -1.375  0.6178159
## group36v(12+24)   -5.125  1.0700884
## group48vall      -16.000  1.5133336
##                    t value     Pr(>|t|)
## group12v24       -2.225582 3.427037e-02
## group36v(12+24)  -4.789324 4.933506e-05
## group48vall     -10.572685 2.775713e-11
## attr(,"class")
## [1] "fit_contrast"
```

This output provides estimate, se(est), t and p-value. It is equal to the lm() output and the aov() output above. If you want a column for degrees of freedom, include the argument df=T. You can add the argument conf.int=.95 to get confidence intervals for the contrast estimates. This command doesn't offer the Welch test though (last time I checked at least) so it only applies when the equality of variance assumption holds.

The fit.contrast function can also handle nonorthogonal contrasts. But if you choose to use the lm() command with nonorthogonal contrasts be careful because the lm() does something different with nonorthogonal contrasts, which will be explained after we do regression. So be careful if you use lm() with non-orthogonal contrasts. While fit.contrast() can allow non-orthogonal contrasts it is limited in that there can't be more than number of groups minus 1 contrasts in a single command, so if you want to test more than number of groups minus 1 non-orthogonal contrasts, then run multiple fit.contrast commands.

Let me illustrate. Here is the fit.contrast command for the first 3 contrasts in Appendix 2. Output from the fit.contrast() command matches the pooled contrast version in SPSS. But the lm() version gives a different

answer.

```
# first 3 non-orthogonal contrasts
fit.contrast(out.model, "group", rbind(cont1 = c(1,
    -1, 0, 0), cont2 = c(1, 0, -1, 0), cont3 = c(1,
    0, 0, -1)))


##            Estimate Std. Error
## groupcont1   -1.375  0.6178159
## groupcont2   -3.250  0.6178159
## groupcont3   -6.875  0.6178159
##                t value     Pr(>|t|)
## groupcont1  -2.225582 3.427037e-02
## groupcont2  -5.260467 1.361059e-05
## groupcont3 -11.127911 8.644137e-12
## attr(,"class")
## [1] "fit_contrast"


# output matches appendix 2 first three contrasts

# lm version
contrasts(sleep.data$group) <- cbind(C1 = c(1, -1,
    0, 0), C2 = c(1, 0, -1, 0), C3 = c(1, 0, 0, -1))
# check the command worked as intended
contrasts(sleep.data$group)


##   C1 C2 C3
## 1  1  1  1
## 2 -1  0  0
## 3  0 -1  0
## 4  0  0 -1


output.lm <- lm(performance ~ group, sleep.data)
summary(output.lm)


##
## Call:
## lm(formula = performance ~ group, data = sleep.data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.3750 -0.6562  0.0000  0.6562  2.3750
##
## Coefficients:
##             Estimate Std. Error t value
## (Intercept)  22.2500     0.2184 101.863
## groupC1       1.5000     0.3783   3.965
```

```
## groupC2      -0.3750     0.3783  -0.991
## groupC3      -4.0000     0.3783 -10.573
##             Pr(>|t|)
## (Intercept)  < 2e-16 ***
## groupC1     0.000462 ***
## groupC2     0.330082
## groupC3     2.78e-11 ***
## ---
## Signif. codes:
##   0 '***' 0.001 '**' 0.01 '*' 0.05
##   '.' 0.1 ' ' 1
##
## Residual standard error: 1.236 on 28 degrees of freedom
## Multiple R-squared:  0.833,Adjusted R-squared:  0.8151
## F-statistic: 46.56 on 3 and 28 DF,  p-value: 5.222e-11


# t.tests are not the same as with SPSS nor
# fit.contrast; yikes
```

These alternative methods can lead to different conclusions. One needs to be careful about what one is doing and make sure the correct commands are being tested; don't just assume you can extend a command if you don't understand what you are doing. Here fit.contrast() produces the correct tests but lm() doesn't unless you do the contrasts correctly.

If you want to know how to get lm() to produce the right output for now you need to apply a different set of contrasts that involves the inverse of the transpose of the contrast matrix (that's matrix algebra terminology that we will cover next term). I illustrate here with the first three contrasts in Appendix 2 that didn't reproduce correctly in the lm() command earlier but now we can get them reproduced correctly.

```
# write the contrast matrix you want note: need
# to augment the contrasts with the unit contrast
contmat <- cbind(c(1, 1, 1, 1), c(1, -1, 0, 0), c(1,
    0, -1, 0), c(1, 0, 0, -1))
contmat


##      [,1] [,2] [,3] [,4]
## [1,]    1    1    1    1
## [2,]    1   -1    0    0
## [3,]    1    0   -1    0
## [4,]    1    0    0   -1


# transform that to a new contrast matrix
newcontmat <- solve(t(contmat))
newcontmat


##      [,1]  [,2]  [,3]  [,4]
```

```
## [1,] 0.25  0.25  0.25  0.25
## [2,] 0.25 -0.75  0.25  0.25
## [3,] 0.25  0.25 -0.75  0.25
## [4,] 0.25  0.25  0.25 -0.75


# attach the new contrast matrix to the factor
# you want to test and drop the unit vector at
# this stage, then rerun lm
contrasts(sleep.data$group) <- newcontmat[, -1]
output.lm <- lm(performance ~ group, sleep.data)
summary(output.lm)


##
## Call:
## lm(formula = performance ~ group, data = sleep.data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.3750 -0.6562  0.0000  0.6562  2.3750
##
## Coefficients:
##             Estimate Std. Error t value
## (Intercept)  22.2500     0.2184 101.863
## group1       -1.3750     0.6178  -2.226
## group2       -3.2500     0.6178  -5.260
## group3       -6.8750     0.6178 -11.128
##             Pr(>|t|)
## (Intercept)  < 2e-16 ***
## group1        0.0343 *
## group2      1.36e-05 ***
## group3      8.64e-12 ***
## ---
## Signif. codes:
##   0 '***' 0.001 '**' 0.01 '*' 0.05
##   '.' 0.1 ' ' 1
##
## Residual standard error: 1.236 on 28 degrees of freedom
## Multiple R-squared:  0.833,Adjusted R-squared:  0.8151
## F-statistic: 46.56 on 3 and 28 DF,  p-value: 5.222e-11
```

Look at the terms in the newcontmat. In order to test the (1, -1, 0, 0) contrast in the lm regression command we need to ask R to test the different contrast (.25, -.75, .25, .25). This will seem weird now, but it will make sense later in the term once we learn regression methods. You may want to use the fit.contrast command for now rather than the lm() command when testing nonorthogonal contrasts until we get to those more complicated concepts.

The explanation for now: we learned contrasts as weights that apply to the cell means to produce the "I hats". But in the lm() setting we are applying the contrasts to the "I hats" in order to recover, or model, the cell

means, so in a sense we need the inverse of the contrast matrix (i.e., the reverse operation) because rather than going from means to "I hats" we are going from "I hats" to means. The inverse(transpose()) operation sets up the analysis so that our new set of contrasts multiplies the Ihats (equivalent to the regression betas) to produce the cell means. The lm command also makes use of the unit vector and so augments the contrast matrix to include the unit vector representing the grand mean; let's refer to this augmented contrast matrix as X. For an orthogonal set X of contrasts that includes the unit vector, the inverse is equal to the transpose so the computation returns the orthogonal matrix of contrasts X and lm() will work correctly with the original set of orthogonal contrasts (technically, the operation inverse(t(X)) returns an orthonormal version of X, but more on this in a later lecture notes). This issue with lm() is only with nonorthogonal contrasts. The solve() command in base R computes the matrix inverse; other functions to compute the matrix inverse include the inv() function in the matlib package. The inverse operation requires a square matrix so that is why the orthogonal contrast matrix needs to be augmented with the unit vector.

The model.tables command is useful. Once you run the aov() command you use the model.tables command to get tables of means or tables of the structural model values.

```
out.model <- aov(performance ~ group, data = sleep.data)
model.tables(out.model, "effects")


## Tables of effects
##
##  group
## group
##      1      2      3      4
## -2.875 -1.500  0.375  4.000


model.tables(out.model, "means")


## Tables of means
## Grand mean
##
## 22.25
##
##  group
## group
##      1      2      3      4
## 19.375 20.750 22.625 26.250
```

If you also want the standard error (the one based on the homogeneity assumption) printed in the output of model.tables(), then add the argument se=T as in.

```
model.tables(out.model, "means", se = T)


## Tables of means
## Grand mean
##
## 22.25
```

```
##
##  group
## group
##      1      2      3      4
## 19.375 20.750 22.625 26.250
##
## Standard errors for differences of means
##          group
##         0.6178
## replic.      8
```

But be careful with model.tables command when you have unequal sample sizes across the groups. The model.tables command uses a different approach to defining effects and the grand mean than the typical one (and one I've been using in this course). In this course, I've defined the grand mean as the mean of the cell means and each cell effect (the αs) is defined as the cell mean minus the mean of the cell means. However, the model.tables command in R defines the grand mean as the mean of all the data and each cell effect is the cell mean minus the mean of all the data. These two methods are equivalent when the cells have the same sample size, but they differ when the cell sizes are different. The approach I've adopted in the lecture notes is consistent with what is called the regression approach and the approach used in model.tables is known as the hierarchical approach. There are different ways of handling unequal sample sizes and we will cover this in more depth in Lecture Notes 5 and again when we cover multiple regression.

Bonferonni in R

To get all possible pairwise mean tests, use the command pairwise.t.test(). It can perform tests with or without the equal variance assumption using pool.sd=T or pool.sd=F, respectively. The command can also do Bonferonni, with p.adjust="bonferroni" or the false discovery rate method with p.adjust="fdr".

```
attach(sleep.data)
pairwise.t.test(performance, group)
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  performance and group
##
##   1       2       3
## 2 0.034   -       -
## 3 4.1e-05 0.010   -
## 4 5.2e-11 5.9e-09 1.0e-05
##
## P value adjustment method: holm
```

```
pairwise.t.test(performance, group, pool.sd = F)
```

```
##
##  Pairwise comparisons using t tests with non-pooled SD
##
## data:  performance and group
```

```
##
##   1       2       3
## 2 0.04309 -       -
## 3 0.00025 0.01792 -
## 4 1.6e-07 3.0e-06 0.00017
##
## P value adjustment method: holm
```

```
pairwise.t.test(performance, group, p.adjust = "bonferroni")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  performance and group
##
##   1       2       3
## 2 0.206   -       -
## 3 8.2e-05 0.031   -
## 4 5.2e-11 7.0e-09 1.6e-05
##
## P value adjustment method: bonferroni
```

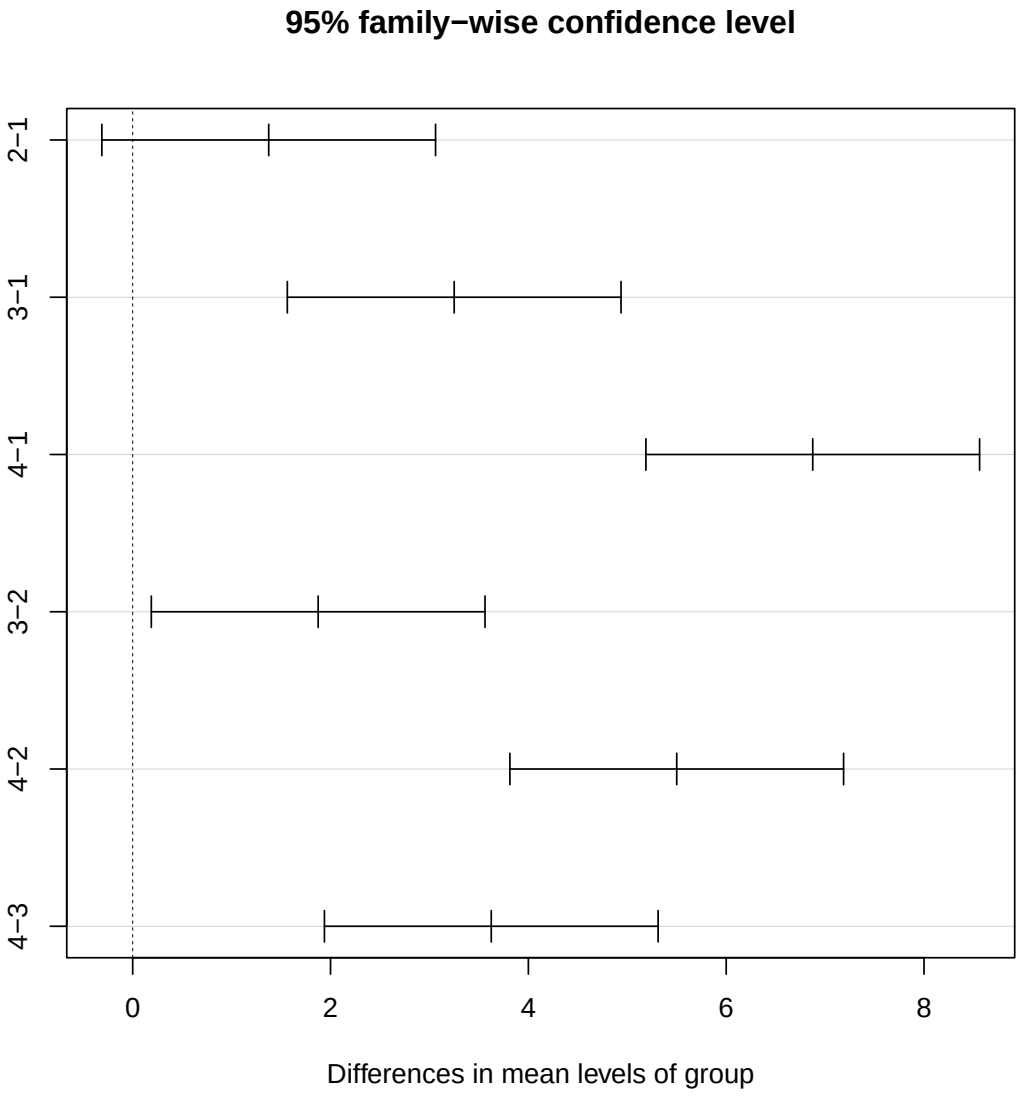
Tukey in R

Tukey tests are done through the TukeyHSD() command, which prints both adjusted Tukey p-value and confidence interval. This syntax chunk assumes we already ran the aov() command and stored the output in out.model.

```
TukeyHSD(out.model, which = "group")
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = performance ~ group, data = sleep.data)
##
## $group
##      diff        lwr     upr     p adj
## 2-1 1.375 -0.3118298 3.06183 0.1409966
## 3-1 3.250  1.5631702 4.93683 0.0000768
## 4-1 6.875  5.1881702 8.56183 0.0000000
## 3-2 1.875  0.1881702 3.56183 0.0250231
## 4-2 5.500  3.8131702 7.18683 0.0000000
## 4-3 3.625  1.9381702 5.31183 0.0000150
```

```
# can also get a useful plot
plot(TukeyHSD(out.model, which = "group"))
```



There are many other libraries and special functions to test contrasts and post hoc comparisons. One package worth examining is multcomp. I'll talk about the emmeans package in Lecture Notes 4, which has some better plotting functions for posthoc tests. There is a function in base R called oneway.test that performs a Welch-type ANOVA on the omnibus F test, but I don't see a way to perform contrasts in the oneway.test framework.

Of course, you can write your own functions in R to organize output the way you want it. That allows you to set up a workflow for commonly used analyses that you perform. For example, here is a function I wrote that computes the Welch separate variance test on contrast to parallel the SPSS ONEWAY separate variance approach:

```
sepvarcontrast <- function(dv, group, contrast) {
    means <- c(by(dv, group, mean))
    vars <- c(by(dv, group, var))
```

```
    ns <- c(by(dv, group, length))
    ihat <- contrast %*% means
    t.denominator <- sqrt(contrast^2 %*% (vars/ns))
    t.welch <- ihat/t.denominator
    num.contrast <- ifelse(is.null(dim(contrast)),
        1, dim(contrast)[1])
    df.welch <- rep(0, num.contrast)
    if (is.null(dim(contrast)))
        contrast <- t(as.matrix(contrast))
    for (i in 1:num.contrast) {
        num <- (contrast[i, ]^2 %*% (vars/ns))^2
        den <- sum((contrast[i, ]^2 * vars/ns)^2/(ns -
            1))
        df.welch[i] <- num/den
    }
    p.welch <- 2 * (1 - pt(abs(t.welch), df.welch))
    result <- list(ihat = ihat, se.ihat = t.denominator,
        t.welch = t.welch, df.welch = df.welch, p.welch = p.welch)
    return(result)
}
```

You call it by giving the function three arguments: dv, group and contrast such as

```
sepvarcontrast(sleep.data$performance, sleep.data$group,
    c(1, -1, 0, 0))
```

```
## $ihat
##         [,1]
## [1,] -1.375
##
## $se.ihat
##           [,1]
## [1,] 0.6178159
##
## $t.welch
##           [,1]
## [1,] -2.225582
##
## $df.welch
## [1] 13.91955
##
## $p.welch
##             [,1]
## [1,] 0.04308851
```

The command can take multiple contrasts in the form of a matrix, such as rbind

```
sepvarcontrast(sleep.data$performance, sleep.data$group,
    rbind(c(1, -1, 0, 0), c(1, 1, -2, 0), c(1, 1, 1,
        -3)))


## $ihat
##         [,1]
## [1,]  -1.375
## [2,]  -5.125
## [3,] -16.000
##
## $se.ihat
##           [,1]
## [1,] 0.6178159
## [2,] 1.0426186
## [3,] 1.5512092
##
## $t.welch
##            [,1]
## [1,]  -2.225582
## [2,]  -4.915508
## [3,] -10.314534
##
## $df.welch
## [1] 13.91955 14.49169 11.51344
##
## $p.welch
##               [,1]
## [1,] 4.308851e-02
## [2,] 2.060439e-04
## [3,] 3.670147e-07
```

This duplicates the output in SPSS for the unequal variance contrast tests (see second example in Appendix 1 of these lecture notes). The separate variance contrast code is not elegant but it works. Future versions of the code can check for missing data, accept formula notation, put the output in the form of table, work with output of either aov() or lm() in the spirit of the fit.contrast() command described above, do Scheffe corrections, and handle factorial designs. Adding such bells and whistles is both the great thing about R but also the bad thing (a huge time commitment to write the code). This command also reproduces a more complicated ANOVA we will handle in LN5 with unequal sample sizes with complicated contrasts like -3,1,-1,3.

In LN11 I'll introduce linear algebra and show how to write code that can test contrasts in very general settings (such as mixed factorial ANOVA designs that have between-subject factors and repeated-measures factors). That may be the general way to handle classic and Welch-type contrasts with and without Scheffe or Bonferroni corrections all in one convenient R function. So more on this in LN11.

Scheffe in R

Here is syntax for Scheffe that reproduces the formulae in the lecture notes. This function assumes you already computed the contrast value $\hat{\text{I}}$ and the standard error of $\hat{\text{I}}$. Those two numbers are output of performing a contrast using the above procedures. Two more arguments are the number of groups and the degrees of freedom for error. The default is $\alpha = .05$ but that can be changed through the alpha argument. For the degrees of freedom and standard error of $\hat{\text{I}}$ you can enter either the traditional values or the Welch values.

```
scheffe <- function(ihat, se.ihat, ngroups, df.error,
    alpha = 0.05) {
    ScheffeFcritical <- (ngroups - 1) * qf(1 - alpha,
        ngroups - 1, df.error)
    Scheffetcritical <- sqrt(ScheffeFcritical)
    S <- Scheffetcritical * se.ihat
    result <- list(Scheffetcritical = Scheffetcritical,
        S = S, lowerCI = ihat - S, upperCI = ihat +
            S)
    return(result)
}
```

This simple Scheffe function prints out the Scheffe t critical, the value S and the confidence interval around S. For an example, double check that you get the same result of Scheffe t critical and S for the psychotherapy/behavioral treatment example. For the first contrast the contrast value was 6, the standard error of the contrast was 2.4698, there were 4 groups and 20 degrees of freedom for the error.

```
scheffe(6, 2.4698, 4, 20)
```

```
## $Scheffetcritical
## [1] 3.048799
##
## $S
## [1] 7.529923
##
## $lowerCI
## [1] -1.529923
##
## $upperCI
## [1] 13.52992
```

The output of this function reproduces the hand computation shown on page 3-42.

Kruskal-Wallis in R

In R the Kruskal-Wallis test is computed through the kruskal.test(dv $\sim$ group) command. See Appendix 4 for relation to ANOVA on ranks.

```
kruskal.test(performance ~ group, data = sleep.data)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  performance by group
## Kruskal-Wallis chi-squared =
## 25.258, df = 3, p-value = 1.364e-05
```

Bayesian Approach to Contrasts

For completeness here is an example for implementing contrasts in a Bayesian framework. Because the command is so similar to the lm() command I'll use the nonorthogonal contrast example I showed earlier with the lm() command where I had to re-specify the contrasts in order to get the correct output (mostly to reinforce that point—had I used a complete set of orthogonal contrasts then lm() and brm() would be correct without needing that extra step).

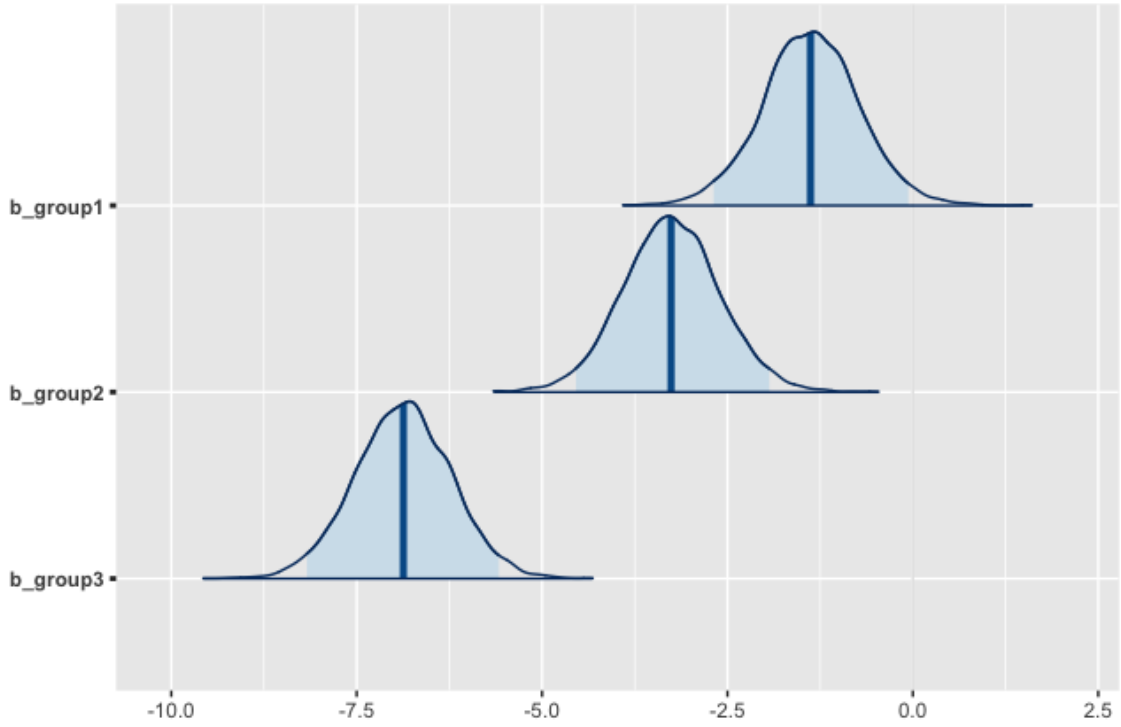
I'll just use the default priors in this example but as shown in Lecture Notes #1 that can be changed. Given that the default priors are not the uniform priors we discussed before the results here will be slightly different than the results from the classical test.

```
# write the contrast matrix you want
contmat <- cbind(c(1, 1, 1, 1), c(1, -1, 0, 0), c(1,
    0, -1, 0), c(1, 0, 0, -1))
# transform that to a new matrix
newcontmat <- solve(t(contmat))
# make the new contrast the one associated with
# the factor you want to test
sleep.data$group <- factor(sleep.data$group)
contrasts(sleep.data$group) <- newcontmat[, -1]

library(brms)
out.bayes <- brm(performance ~ group, data = sleep.data,
    iter = 20000, thin = 5)
summary(out.bayes)
plot(out.bayes)

# can also do a plot with 95% shaded; see also
# the tidybayes package
library(bayesplot)
mcmc_areas(as.matrix(out.bayes), regex_pars = "group",
    prob = 0.95) + yaxis_text(c("c1", "c2", "c3"))
```

The snippets of the output include

```
 Family: gaussian
  Links: mu = identity; sigma = identity
Formula: performance ~ group
   Data: sleep.data (Number of observations: 32)
Samples: 4 chains, each with iter = 20000; warmup = 10000; thin = 5;
         total post-warmup samples = 8000

Population-Level Effects:
          Estimate Est.Error l-95% CI u-95% CI Eff.Sample Rhat
Intercept    22.25      0.23    21.78    22.70       7721 1.00
group1       -1.38      0.66    -2.69    -0.06       7982 1.00
group2       -3.26      0.66    -4.54    -1.94       7660 1.00
group3       -6.88      0.65    -8.17    -5.58       7744 1.00
```

Figure 3-4: Density plots for each of the three contrasts with shaded 95% interval. Even though labels say group1, group2, group3, these correspond to contrasts 1, 2 and 3 respectively.




```
Family Specific Parameters:
      Estimate Est.Error l-95% CI u-95% CI Eff.Sample Rhat
sigma     1.29      0.18     1.00     1.70       8000 1.00
```

Note that the three estimates of the contrast values are similar to those presented above using fit.contrast and the correct lm(), and the conclusions from the 95% confidence intervals are consistent across the various approaches. The density plots are shown in the figure.

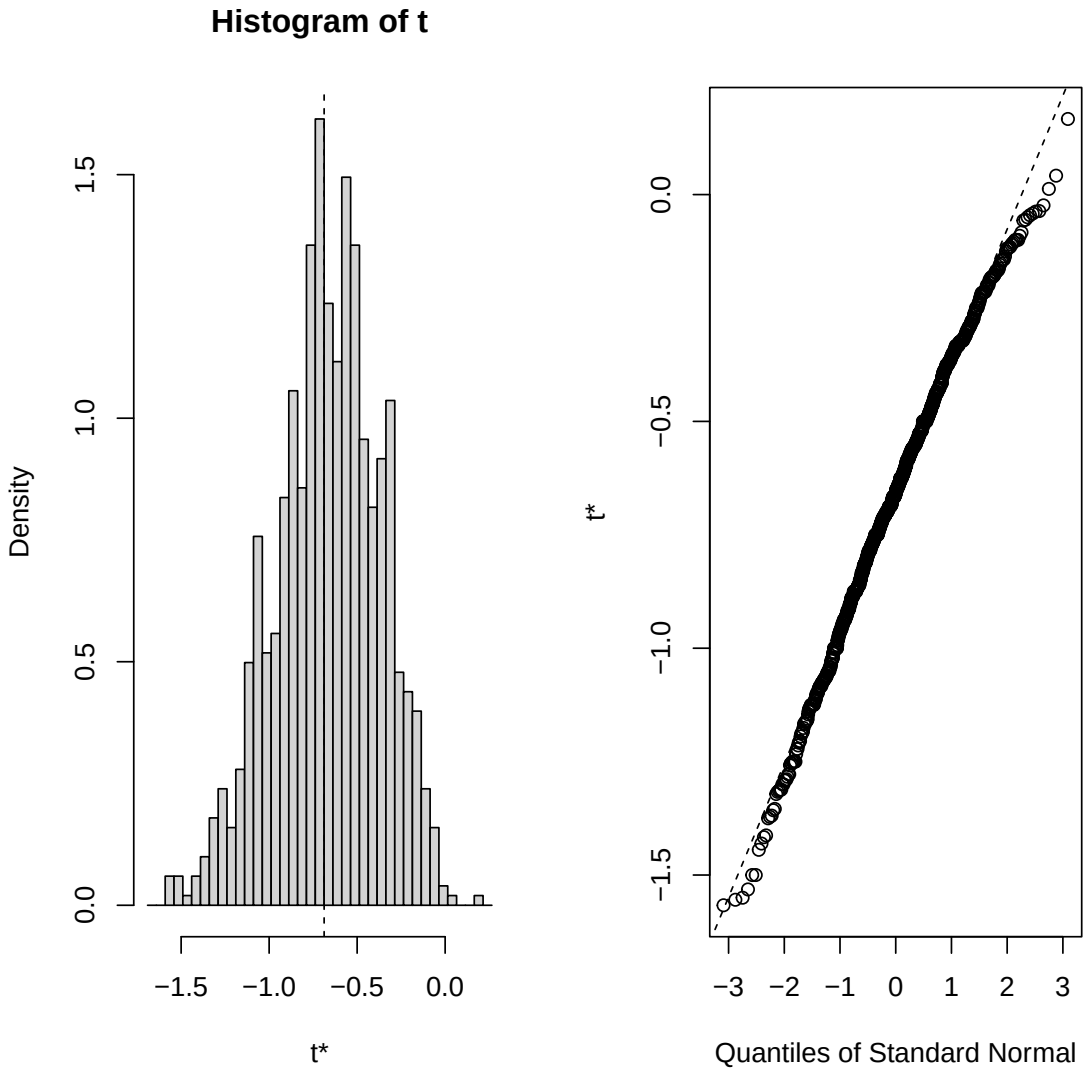
Bootstrap Approach to Contrasts

LN1 briefly touched on bootstrapping for two groups. Here do the analagous computation for contrasts using the nonparametric version of the boostrap that resamples the sample (as mentioned before there are other forms of bootstrap such as resampling the residuals). I'll redo the sleep deprivation ANOVA with a set of orthogonal contrasts using the lm command. Basically, I use lm to compute the "i hats" (the coef command saves the estimates) and boostrap the "i hats".

```
# repeat contrasts for clarity
contrasts(sleep.data$group) <- cbind(C12.24 = c(1,
```

```
    -1, 0, 0), C1224.36 = c(1, 1, -2, 0), C122436.48 = c(1,
    1, 1, -3))
output.lm <- lm(performance ~ group, sleep.data)
summary(output.lm)


##
## Call:
## lm(formula = performance ~ group, data = sleep.data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.3750 -0.6562  0.0000  0.6562  2.3750
##
## Coefficients:
##                 Estimate Std. Error
## (Intercept)      22.2500     0.2184
## groupC12.24      -0.6875     0.3089
## groupC1224.36    -0.8542     0.1783
## groupC122436.48  -1.3333     0.1261
##                 t value Pr(>|t|)
## (Intercept)     101.863  < 2e-16 ***
## groupC12.24      -2.226   0.0343 *
## groupC1224.36    -4.789 4.93e-05 ***
## groupC122436.48 -10.573 2.78e-11 ***
## ---
## Signif. codes:
##   0 '***' 0.001 '**' 0.01 '*' 0.05
##   '.' 0.1 ' ' 1
##
## Residual standard error: 1.236 on 28 degrees of freedom
## Multiple R-squared:  0.833,Adjusted R-squared:  0.8151
## F-statistic: 46.56 on 3 and 28 DF,  p-value: 5.222e-11


# now bootstrap the puppy
library(boot)
mybootfunction <- function(formula, data, indices) {
    d <- data[indices, ]
    fit <- lm(formula, data = d)
    # to avoid confusion I drop the intercept
    return(coef(fit)[-1])
}

# set seed so always get same result
set.seed(14813)
bootresults <- boot(sleep.data, statistic = mybootfunction,
    R = 1000, formula = performance ~ group)
bootresults


##
```

```
## ORDINARY NONPARAMETRIC BOOTSTRAP
##
##
## Call:
## boot(data = sleep.data, statistic = mybootfunction, R = 1000,
##     formula = performance ~ group)
##
##
## Bootstrap Statistics :
##       original       bias    std. error
## t1* -0.6875000  0.022236051   0.2944298
## t2* -0.8541667  0.004432259   0.1630102
## t3* -1.3333333 -0.003929984   0.1264101


# need to index each contrast; here just show
# contrast 1
boot.ci(bootresults, index = 1)


## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 1000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = bootresults, index = 1)
##
## Intervals :
## Level      Normal              Basic
## 95%   (-1.2868, -0.1327 )   (-1.2385, -0.0863 )
##
## Level     Percentile            BCa
## 95%   (-1.2887, -0.1365 )   (-1.3741, -0.1955 )
## Calculations and Intervals on Original Scale


plot(bootresults, index = 1)
```



The 95% confidence interval around "i hat" does not include 0 so we reject the null hypothesis that population "i" is 0.

Appendix 4

The Kruskal-Wallis test is a nonparametric test that can be usaed on between-subjects ANOVA. The usual derivation is very different than ANOVA, using Chi-square rather than F, no obvious sum of squares table, etc., and not having a clear direction for how to perform additional tests such as contrasts, Tukey, and Scheffe. In this Appendix, I'll show an interesting relation between ANOVA on the rank data (i.e., a transformation of the original data to ranks and then perform a regular ANOVA on the ranks) and the Kruskal-Wallis test. This connection, along with many others between standard tests on rank transformed data and nonparametric tests, is explained well in a book by Conover.

```
KRUSKAL-WALLIS TEST ON THE SLEEP DEPRIVATION DATA

SPSS SYNTAX AND OUTPUT:

NPAR TESTS
   /K-W=dv   BY codes(1, 4).

      MEAN RANK    CASES
           6.31        8   CODES =    1   12hr
          11.88        8   CODES =    2   24hr
          19.50        8   CODES =    3   36hr
          28.31        8   CODES =    4   48hr
                      32   TOTAL

                                                    CORRECTED FOR TIES
          CASES    CHI-SQUARE  SIGNIFICANCE    CHI-SQUARE  SIGNIFICANCE
             32       24.8828        0.0000       25.2578        0.0000
```

I showed the identical 25.2578 value in R in the previous Appendix.

Now I'll show the connection between the Kruskal-Wallis test and an ANOVA on rank data.

```
DATA TRANSFORMED INTO RANKS
   12hr   24hr   36hr   48hr
    8.5   13.5   25.5   27.5
    8.5    8.5   21.5   30.0
    1.0   13.5   17.5   24.0
    3.5   17.5   21.5   30.0
    8.5    8.5   13.5   25.5
    3.5    8.5   17.5   32.0
   13.5   21.5   17.5   27.5
    3.5    3.5   21.5   30.0

ONE WAY ANOVA ON THE RANKS
                                  SUM OF         MEAN             F      F
         SOURCE           D.F.    SQUARES       SQUARES          RATIO  PROB.
 BETWEEN GROUPS              3     2189.6875      729.8958      41.0538  .0000
```

```
WITHIN GROUPS             28      497.8125       17.7790
TOTAL                     31     2687.5000
```

The F on ranked data is related to the Kruskal-Wallis χ^2 test (corrected for ties) by this formula

$$\begin{aligned} F_{ranks} &= \frac{H/(T\text{ - }1)}{(N\text{ - }1\text{ - }H)/(N\text{ - }T)} \\ 41.05 &= \frac{25.258/3}{(31-25.258)/28} \end{aligned}$$

where H is the Kruskal-Wallis χ^2 result (corrected for ties), T is the number of groups and N is the total number of subjects. Thus, the only ingredients that relate K-W's H to the ANOVA's F on ranked data are the number of groups (T) and the number of subjects (N). For relatively large N, the *p* value from the two tests will be indistinguishable.

In R we can perform the rank transformation easily and continue our usual ANOVA analytic workflow such as performing contrasts, Tukeys, etc., but using the ANOVA on ranks rather than the ANOVA on the original data. The interpretation of these results is based on the average ranks, such as a significant contrast means that the "Ihat" on the average ranks is statistically significant, the significant Tukey test is interpreted as a difference of two pairwise mean ranks, etc. The benefit of using the rank transformation is that the analysis is not sensitive to outliers in the sense that if you change the values of the outliers without changing the ranks, the analysis remains identical; whereas, for an ANOVA on raw data, if the value of an outlier changes without affecting the ranks, the means could still change dramatically. Basically, means on raw data are more sensitive to outliers, nonnormal distributions, etc., than means on rank data. The logic of nonparametric tests is to take advantage of this idea and develop a parallel universe of statistical tests on the rank data. Nonparametric tests still make assumptions, such as independence and that data are sampled from a common distribution (just not a requirement that the common distribution be a normal distribution), so nonparametric tests are not completely assumption free as you might read on the web.

We can conceptualize the act of converting raw data to ranks as a transformation of the raw data to a new scale analogous to what we learned in LN2 surrounding power transformations.

```
sleep.data$rank.performance <- rank(sleep.data$performance)
out.aov.rank <- aov(rank.performance ~ group, data = sleep.data)
summary(out.aov.rank)


##             Df Sum Sq Mean Sq F value
## group        3 2189.7   729.9   41.05
## Residuals   28  497.8    17.8
##               Pr(>F)
## group       2.21e-10 ***
## Residuals
## ---
## Signif. codes:
##   0 '***' 0.001 '**' 0.01 '*' 0.05
##   '.' 0.1 ' ' 1


# for good measure, we can test a contrasts on
```

```r
# the rank data; could then apply scheffe use use
# the welch-test for the contrast, or followup
# with Tukey, etc
fit.contrast(out.aov.rank, "group", rbind(cont1 = c(1,
    -1, 0, 0), cont2 = c(1, 0, -1, 0), cont3 = c(1,
    0, 0, -1)))


##            Estimate Std. Error
## groupcont1  -5.5625   2.108259
## groupcont2 -13.1875   2.108259
## groupcont3 -22.0000   2.108259
##               t value     Pr(>|t|)
## groupcont1  -2.638433 1.344870e-02
## groupcont2  -6.255162 9.236360e-07
## groupcont3 -10.435152 3.727567e-11
## attr(,"class")
## [1] "fit_contrast"
```