

Keppel, G. & Wickens, T. D. *Design and Analysis*
Chapter 4: Analytical Comparisons Among Treatment Means

4.1 The Need for Analytical Comparisons

- “...the between-groups sum of squares *averages* the differences between the different pairs of means. As such, it cannot provide unambiguous information about the exact nature of treatment effects.”
- “Because SS_A is a composite, an F ratio based on more than two treatment levels is known as an *omnibus* or *overall* test.” That is, with two levels we can say which condition is larger, but with more than two levels, we don’t know which conditions are actually different (based solely on the F ratio). In order to know which particular means differ, we have to compute specific comparisons between means.
- When computing SS_A , we’re actually computing how far each condition mean is from every other condition mean. Thus, as seen in Formula 4.1, SS_A is actually the average squared distance among the group means. To illustrate for K&W51 (what else!):

$$SS_A = [A] - [T] = 3314.25$$

$$SS_A = n \sum (\bar{Y}_A - \bar{Y}_T)^2 = 4 \sum (\bar{Y}_A - 45.875)^2 = 4 (375.4 + 66 + 135.1 + 252) = 3314$$

$$SS_A = \frac{n}{a} \sum_{pairs} (\bar{Y}_j - \bar{Y}_k)^2 =$$

(4.1)

$$4/4 [(26.5 - 61.75)^2 + (37.75 - 61.75)^2 + (57.5 - 61.75)^2 + (26.5 - 57.5)^2 + (37.75 - 57.5)^2 + (26.5 - 37.75)^2] = 3314.25$$

- If our ANOVA tells us that we have insufficient evidence to reject H_0 (essentially a Type II error), then we would consider ways to make our study more powerful or we might move along to a different study entirely. However, in the presence of a significant F ratio (and more than two conditions) we would need to conduct additional analyses to disambiguate the results of our analyses. For that purpose, we would need to compute post hoc comparisons.
- Ideally, we might avoid the omnibus test entirely and conduct a series of specific planned comparisons. The reality we must face, however, is that some people (e.g. journal editors) may not trust that our comparisons were truly planned.
- Regardless of the type of comparison (planned or post hoc) you may be considering, the actual computation of the comparison is the same.

4.2 An Example of Planned Comparisons

- K&W provide a clear explanation of the value of planned comparisons using McGovern (1964). Her experiment included five treatment conditions in which a second task (after learning a vocabulary list) might interfere with memory for the originally learned vocabulary items. Particular treatments were thought to interfere with particular

components of memory (which K&W generically label A, B, and C). One treatment condition (a_1) was a control condition that should not interfere with any memory component. Another treatment condition (a_2) was thought to interfere with Component A, but not B and C. These hypotheses could actually be translated into a series of planned comparisons.

- “The McGovern study demonstrates the analytical possibilities of experimentation— that is, creating an experimental design that yields several focused comparisons that are central to the purposes of the experiment. Quite obviously, these are *planned* comparisons that would be tested directly.”

4.3 Comparisons Between Treatment Means

- First, for terminology, K&W use *comparisons*, *contrasts*, and *single-df comparisons* interchangeably. (Because you’re always comparing two means, the $df_{Comp} = 1$.)
- A typical comparison is between two means, but that doesn’t imply that you can only compare two groups at a time (a simple pairwise comparison). Your null hypothesis for such a pairwise comparison would look like this:

$$H_0: \mu_1 = \mu_2 \text{ OR } \mu_1 - \mu_2 = 0$$

- It is perfectly reasonable to compare combined groups (a complex comparison). Thus, the null hypothesis for a comparison of one group with the combined scores from two other groups would be:

$$H_0: \mu_1 = \frac{\mu_2 + \mu_3}{2} \text{ OR } \mu_1 - \frac{\mu_2 + \mu_3}{2} = 0$$

Because it’s not always clear that a complex comparison is reasonable, K&W admonish us to scrutinize any complex comparison. Exactly why are the groups being combined?

- The symbol ψ represents the comparison, as in:

$$\psi_1 = \mu_1 - \frac{\mu_2 + \mu_3}{2} \text{ OR } \psi_1 = (+1)(\mu_1) + (-.5)(\mu_2) + (-.5)(\mu_3)$$

Note the equivalence of the two statements above. The second expression, however, is useful because it introduces the notion of coefficients. Each of the means is multiplied by a coefficient prior to adding the products to get the value of ψ . In general:

$$\psi = \sum(c_j)(\mu_j) \text{ and } \sum(c_j) = 0 \tag{4.2 and 4.3}$$

- For example, in an experiment with 5 levels (e.g., see Table 4.1) you might want to compare groups 2 and 4 with groups 1, 3, and 5, so your complex comparison might look like:

$$\psi = (+2)(\mu_1) + (-3)(\mu_2) + (+2)(\mu_3) + (-3)(\mu_4) + (+2)(\mu_5)$$

Note that I've avoided fractions by using whole number weights. Doing so seems easier to me, but note what K&W say about the standard form of coefficients (p. 68).

- Because we've been dealing with Greek symbols, you should recognize that we're going to have to do some estimation. Thus:

$$\hat{\psi} = \sum (c_j)(\bar{Y}_j) \quad (4.4)$$

- OK, let's take a step back and see where we're heading. Ultimately, we're going to compute an F ratio for a comparison. To create the $F_{Comparison}$, we're going to need an $MS_{Comparison}$ (MS_v) and an MS_{Error} . What we're doing now is trying to get the $MS_{Comparison}$ by first computing the $SS_{Comparison}$. The formulas that will give us what we want are:

$$SS_{Comp} = \frac{n(\hat{\psi})^2}{\sum c_j^2} \quad \text{OR, by substitution, } SS_{Comp} = \frac{n[\sum (c_j)(\bar{Y}_j)]^2}{\sum c_j^2} \quad \text{OR}$$

$$SS_{Comp} = \frac{[\sum (c_j)(A_j)]^2}{n[\sum c_j^2]}$$

- The MS_{Comp} is identical to the SS_{Comp} , because the $df = 1$, right? So, all that's left to do to compute the F_{Comp} is to determine the appropriate MS_{Error} for the denominator of the F ratio. The appropriate error term is more difficult to determine when you have some concerns about heterogeneity of variance. For the time being, let's assume that we have no concerns about heterogeneity of variance, which means that we could comfortably use the estimate of σ^2 from the overall ANOVA in our comparisons ($MS_{S/A}$).
- Let's use K&W51 to compute examples of a pairwise comparison and a complex comparison. First, let's compute a simple pairwise comparison of the 4Hour Deprivation group and the 20Hour Deprivation group. The coefficients would be $\{+1, 0, -1, 0\}$. Using the formulas we'd get:

$$SS_{Comp} = \frac{4(26.50 - 57.50)^2}{2} = 1922 \quad \text{OR} \quad SS_{Comp} = \frac{(106 - 230)^2}{4(2)} = 1922$$

Note that the actual weights you choose to use don't matter.

$$SS_{Comp} = \frac{4(53 - 115)^2}{8} = 1922, \text{ using weights } \{+2, 0, -2, 0\}$$

$$SS_{Comp} = \frac{4(5300 - 11500)^2}{80000} = 1922, \text{ using weights } \{+200, 0, -200, 0\}$$

Therefore, $F_{Comp} = \frac{1922}{150.46} = 12.77$

[Note that $MS_{Comp} = 1922$, because $df_{Comp} = 1$. MS_{Error} is from the overall ANOVA.]

Compute a simple comparison of the 4Hour Deprivation group and the 28Hour Deprivation group.

• Now, let's compute a complex comparison of the 4Hour and the 12Hour Dep groups together against the 20Hour and the 28Hour Dep groups together. The coefficients could be $\{+1, +1, -1, -1\}$ or $\{-.5, -.5, .5, .5\}$. The SS_{Comp} and F_{Comp} would be:

$$SS_{Comp} = \frac{[(106 + 151) - (230 + 247)]^2}{4(4)} = \frac{-220^2}{16} = 3025$$

$$F_{Comp} = \frac{3025}{150.46} = 20.11$$

Note, again, that the actual weights don't matter. Using weights $\{+10, +10, -10, -10\}$:

$$SS_{Comp} = \frac{[(1060 + 1510) - (2300 + 2470)]^2}{4(400)} = \frac{-2200^2}{1600} = 3025.$$

Compute a complex comparison of the 4Hour Deprivation group versus the 20Hour and 28Hour Deprivation groups together.

As planned comparisons (among a group of $a - 1$ or fewer comparisons), you'd evaluate these comparisons relative to $F_{Crit}(1,12) = 4.75$.

• Note, however, that computing SS_{Comp} is identical to computing $SS_{Treatment}$ (SS_A), except that only two groups are involved. That is, you can use the exact same formulas that you've already learned, but you simply apply them to the two groups involved in your comparison. Thus,

$$SS_{Comp1vs3} = [A] - [T] = \frac{106^2 + 230^2}{4} - \frac{(106 + 230)^2}{8} = 1922$$

$$SS_{Comp1+2vs3+4} = \frac{(106 + 151)^2 + (230 + 247)^2}{8} - \frac{734^2}{16} = 3025$$

Use the bracket terms to compute SS for the simple and complex comparisons you'd computed earlier using the coefficient weight formula.

- In essence, then, you could avoid learning any new formulas for comparisons, as long as you recognize that all that's involved is a computation of a $SS_{Treatment}$. A further implication is that you can simply have a statistical package compute an ANOVA on the two groups involved in the comparison and the $SS_{Treatment}$ that gets printed out is the $SS_{Comparison}$. And, of course, because the df for a comparison is 1, $SS_{Comparison} = MS_{Comparison}$. If you weren't concerned about heterogeneity of variance, then you'd compute the F_{Comp} by dividing the MS_{Comp} by the $MS_{S/A}$ from the overall ANOVA (as was shown above).

4.4 Evaluating Contrasts with a t Test

- When SPSS, for example, computes a test of a contrast, it reports the results in terms of a t statistic and not F . That said, of course, for two means there is no difference in the interpretation of a t or an F .
- The interpretations are the same because $t^2 = F$. So, by squaring t and tidying up a bit, note what happens to the numerator of the t test. Because $a = 2$ (for a comparison), you'll see in the formula below that the numerator of the t test is the same as SS_A and because $df = 1$, MS_A would be the same. The denominator of the t -test is the pooled variance estimate (which is the same as $MS_{S/A}$). So, the squared t -test is really identical to the ANOVA. Neat, huh?

$$t^2 = \left[\frac{(\bar{Y}_{A_1} - \bar{Y}_{A_2})}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}} \right]^2 = \frac{(\bar{Y}_{A_1} - \bar{Y}_{A_2})^2}{\frac{2s_p^2}{n}} = \frac{n(\bar{Y}_{A_1} - \bar{Y}_{A_2})^2}{s_p^2}$$

- Another way to convince yourself of the relationship between t and F is to look at the tables of critical values for the two statistics. In Table A.1 (for F), look up $F_{Crit}(1,10)$ for $\alpha = .05$. Then, in Table A.2 (for t), look up t_{Crit} for $df = 10$. You should note that $t^2 = F$.
- K&W discuss the possibility of conducting directional comparisons (pp. 73-75). However, my advice is to eschew directional tests (as well as obfuscation ☺). K&W mention some reasons that directional tests might be problematic, to which I'd add the

cynical editor who might well wonder if you really had a directional hypothesis in the first place.

- K&W illustrate how you can compute confidence intervals for a contrast when you compute t .

4.5 Orthogonal Contrasts

- “The valuable property of orthogonal comparisons is that they reflect *independent* or *nonoverlapping* pieces of information.” Thus, knowing the outcome of one comparison gives no indication whatsoever about the outcome of another comparison, if the two are orthogonal.

- Two comparisons are orthogonal if the sum of the products of the two sets of coefficients is zero. Thus, a test for orthogonality of two comparisons is:

$$\sum (c_{1j})(c_{2j}) = 0 \tag{4.20}$$

- “If all contrasts in a set are orthogonal to one another, they are said to be mutually orthogonal.”

- “There can be no more mutually orthogonal contrasts than the degrees of freedom associated with the omnibus effect.”

- That is, with a treatment means, there are only $a-1$ comparisons that are orthogonal to each other and to \bar{Y}_T . Because they represent nonoverlapping information, the sum of the SS_{Comp} for the $a - 1$ orthogonal comparisons is SS_A . “In this sense, the sum of squares obtained from a set of a treatment means is a *composite* of the sums of squares associated with $a - 1$ mutually orthogonal contrasts.”

- For K&W51, one set of orthogonal comparisons would be:

Comp1:	1	0	-1	0
Comp2:	0	1	0	-1
Comp3:	1	-1	1	-1

Test for Orthogonality:

Comp1 vs. Comp2:	(1)(0) + (0)(1) + (-1)(0) + (0)(-1)	= 0
Comp1 vs. Comp3:	(1)(1) + (0)(-1) + (-1)(1) + (0)(-1)	= 0
Comp2 vs. Comp3:	(0)(1) + (1)(-1) + (0)(-1) + (-1)(-1)	= 0

Furthermore, note that $SS_{Comp1} = 1922$, $SS_{Comp2} = 1152$, $SS_{Comp3} = 240.25$. If we add all three of these SS we get 3314.25, which is SS_A from the overall ANOVA. Note that a fourth comparison would, by definition, not be orthogonal (and would lead to a sum of the SS that would necessarily be greater than SS_A).

- Orthogonality is an important property of comparisons. However, your comparisons should be driven by theoretical issues rather than by consideration of orthogonality. (Note K&W’s comments on the McGovern study, including “You should strive to look at

orthogonal comparisons whenever you can, but not to the extent of ignoring substantively important comparisons.”)

4.6 Composite Contrasts Derived from Theory

- “The most sensitive contrast to a pattern of means is the one whose coefficients reflect the same pattern.” That is, suppose that you are conducting a study in which you have some reason (e.g., based on past research) to think that the means will fall in a certain pattern. You could then construct a set of coefficients as follows:

1. Obtain your set of predicted means and use them as coefficients. Suppose, for example, that for K&W51, we had reason to believe that the pattern of means would be: 26.5, 37.75, 57.5, and 61.75. (What an amazing prediction, eh?) These coefficients would clearly not sum to zero: {26.5, 37.75, 57.5, 61.75}.

2. Subtract the grand mean (45.875) from each of the predicted means. That will produce a set of coefficients that does sum to zero: {-19.375, -8.125, 11.625, 15.875}.

3. Those coefficients are fairly messy, so to simplify them I could round a bit {-19.4, -8.1, 11.6, 15.9}. I could also work to try to turn these coefficients into integers.

- The point is that if your predictions about the means are accurate, you will find that the SS_{Comp} for this one comparison (that involves all 4 means) will be very close to the SS_A for an overall ANOVA. To the extent that your prediction is not accurate, then your SS_{Comp} will not “capture” the same variability as that found in the SS_A . For the K&W51 data set, you’d obtain:

$$\hat{\psi} = (-19.4 \times 26.5) + (-8.1 \times 37.75) + (11.6 \times 57.5) + (15.9 \times 61.75) = 828.95$$

$$SS_{\psi} = \frac{4(828.95)^2}{(-19.4^2) + (-8.1^2) + (11.625^2) + (15.9^2)} = \frac{2748632.41}{829.34} = 3314.24$$

The SS_A from the original analysis was 3314.25, so you can see that you’ve “captured” virtually all of the variability in the $SS_{Treatment}$ with just this one contrast.

- As a means of testing theoretical predictions, K&W build on the work of others (e.g., Levin & Neumann, 1999; Rosnow & Rosenthal, 1995). They suggest the construction of a contrast set that assesses the fit of a theoretical prediction with the obtained data, followed by a test of the residual variability (called $SS_{Failure}$). This lack-of-fit test should make some sense to you. That is, you may first establish that you have a reasonable theoretical prediction (akin to finding a significant correlation between two variables). However, it may be that the residual variability (akin to $1 - r^2$) is still quite high. For the example above, obviously, there would be virtually no $SS_{Failure}$.

4.7 Comparing Three or More Means

- For experiments with several levels of the IV, it may make sense to conduct analyses of subsets of the conditions. K&W provide an example of an experiment with a control

condition and several treatment conditions. For that experiment, it may make sense to first test all the treatment conditions to see if they differ. If they don't, then you could possibly combine them in a comparison with the control condition. On the other hand, if any of the treatment conditions differed, you could then compare them individually with the control condition.