

Keppel, G. & Wickens, T. D. *Design and Analysis*
Chapter 6: Simultaneous Comparisons and the Control of Type I Errors

- You should design your research with specific questions in mind, which you then test with specific analyses. However, your design will often lend itself to additional analyses, which may well allow you to learn more about the operation of the variables in question. These additional analyses come with the “burden” that you’ll have a greater chance of making a Type I error among the additional analyses. This chapter covers means of controlling Type I error among these simultaneous comparisons.

6.1 Research Questions and Type I Error

- The family of tests is a set of tests you intend to compute to address a set of research questions. The familywise Type I error rate (α_{FW}) is the probability of making at least one Type I error in the family of tests when all H_0 are true.
- When you consider the huge set of possible post hoc tests one might compute, then you are considering the experimentwise error rate (α_{EW}). Needless to say, it will typically be the case that $\alpha_{FW} < \alpha_{EW}$.
- With a per-comparison error rate (α), you can compute the familywise Type I error rate for a number of comparisons (c) as:

$$\alpha_{FW} = 1 - (1 - \alpha)^c \tag{6.1}$$

- Thus, using K&W51 as an example, if you intended to compute three comparisons using $\alpha = .05$ for each comparison, your α_{FW} would be .14. Though not as accurate as the above formula, for a quick and dirty estimate of α_{FW} you could simply use $c\alpha$ (which in this case would give you an estimate of .15). Of course, as the number of comparisons grows, so does α_{FW} . To convince yourself of this relationship between c and α_{FW} , compute α_{FW} for the number of comparisons indicated below:

| Number of Comparisons (c) | α_{FW} |
|-------------------------------|---------------|
| 4 | |
| 5 | |
| 6 | |

- The formula for computing α_{FW} only works for orthogonal comparisons (i.e., assumed independence), but α_{FW} also increases with an increasing number of nonorthogonal comparisons. Thus, because for K&W51 there are only 3 orthogonal comparisons, the estimates of α_{FW} above are not accurate, though they still make the point that α_{FW} will increase with increasing numbers of comparisons.
- Decreasing your per-comparison error rate (α) will also serve to decrease your α_{FW} .
- K&W distinguish the types of questions that experimenters might ask of their data. If the questions are the relatively small number (e.g., rarely more than $a - 1$) of primary questions,

K&W suggest that no adjustment of the per-comparison error rate (α) is necessary. I'm not sure that journal editors would agree with this suggestion.

- It's much more typical that one would want to conduct a set of comparisons computed to understand the omnibus ANOVA. Sometimes the number of these comparisons is quite limited. Sometimes you want to compute all possible simple pairwise comparisons. And sometimes you may be interested in exploring a fairly large set of simple and complex comparisons. The approach for controlling α_{FW} varies for the different situations.
- Especially when the number of tests might be fairly large, it makes sense to adopt an α_{FW} that is greater than .05 (e.g., .10). Keep in mind, however, that the guidelines for choosing α_{FW} , or choosing a strategy for controlling Type I error in such simultaneous comparisons are not rigid and universally agreed on.

6.2 Planned Comparisons

- OK, we'll discuss planned comparisons, but keep in mind that journal editors might not trust that you've actually planned the comparisons in advance. My advice would be to treat all comparisons as post hoc—at least until you've achieved the sort of stature in the discipline that buys you some slack from editors. ☺
- Planned comparisons must be specified in the initial design of an experiment. They are essential and pivotal tests—not a “plan” to conduct a “fishing expedition.”
- “For clearly planned tests, familywise error correction is generally deemed unnecessary.”
- The comparisons that you choose to compute should be driven by theoretical concerns, rather than concerns about orthogonality. However, orthogonal comparisons should be used because they “keep hypotheses logically and conceptually separate, making them easier to interpret.” Nonorthogonal comparisons must be interpreted with care because of the difficulty of making inferences (interpreting the outcomes). For instance, in an earlier edition, Keppel wondered, “If we reject the null hypothesis for two nonorthogonal comparisons, which comparison represents the ‘true’ reason for the observed differences?”
- Don't allow yourself to be tempted into computing a large number of planned comparisons. For an experiment with a levels, there are $1 + ((3^a - 1) / 2) - 2^a$ comparisons (simple pairwise *and* complex) possible. *Please*, don't ever compute all possible comparisons! If you think carefully about your research, a much smaller set of planned comparisons would be reasonable.
- A common suggestion for multiple planned comparisons is to conduct up to $a - 1$ comparisons with each comparison conducted at $\alpha = .05$. The implication of this suggestion is that people are willing to tolerate a familywise error rate of $(a - 1)(\alpha)$. Thus, in an experiment with 5 levels, you could comfortably compute 4 planned comparisons with each comparison tested using $\alpha = .05$, for a familywise error rate of $\sim .19$. As the number of planned comparisons becomes larger than $a - 1$, consider using a correction for α_{FW} (e.g., Sidák-Bonferroni procedure).

6.3 Restricted Sets of Contrasts

The Bonferroni Procedure

- “The most widely applicable familywise control procedure for small families is the Bonferroni correction.”
- The Bonferroni inequality ($\alpha_{FW} < c \alpha$) states, “The familywise error rate is always less than the sum of the per-comparison error rates of the individual tests.”
- Thus, to ensure that our α_{FW} is kept to a certain level, we could choose that level (e.g. .05 or .10 depending on our preference) and then divide that value by the number of comparisons we wish to compute. Assuming that you are comfortable with $\alpha_{FW} = .10$ and you are about to compute 5 comparisons, you would treat comparisons as significant if they occur with $p \leq \alpha$ (the per-comparison rate), which would be .02 here. Given that SPSS prints out a t and its associated p -value when you ask it to compute contrasts, you’d be able to assess the significance of the t statistic by comparing it to your Bonferroni per-comparison rate.
- For hand computation of such tests, you occasionally need to compute a critical value of t or F for a per-comparison error rate that is not found in the table of critical values of F (A.1). For example, using K&W51, suppose that you wanted to compute 6 simple pairwise comparisons (e.g., 4 vs. 12, 4 vs. 20, etc.). If you were comfortable with $\alpha_{FW} = .10$, your per-comparison error rate (α) would be .0167. (If you preferred $\alpha_{FW} = .05$, your per-comparison error rate would be .008.) Let’s presume that we’re using $\alpha = .0167$. Given homogeneity of variance (a topic that arises in Ch. 7) for K&W51, we would use the overall error term ($MS_{S/A} = 150.458$) for any comparison. Thus, $df_{Error} = 12$ and $df_{Comparison}$ is always 1. In Table A.1, for those df we see the following tabled α values:

| α | F_{Crit} |
|----------|------------|
| .100 | 3.18 |
| .050 | 4.75 |
| .025 | 6.55 |
| .010 | 9.33 |
| .001 | 18.64 |

Although our α is not tabled, we can see that F_{Crit} for our α would be less than 9.33 and greater than 6.55. (And, of course, we could determine t_{Crit} values by taking the square root of the tabled F_{Crit} values.) Suppose that you compute $F_{Comparison} = 10.0$ (or any $F \geq 9.33$). You would conclude that the two groups came from populations with different means (reject H_0). Suppose, instead, that you compute $F_{Comparison} = 6.0$ (or any $F \leq 6.55$). You would conclude that you had insufficient evidence to claim that the two groups came from populations with different means (retain H_0). The tricky stage arises if your $F_{Comparison} = 9.0$. To assess the significance of this outcome, you need to actually compute F_{Crit} for $\alpha = .0167$. You can always use the formula below to determine the F for a given level of α .

$$t = z + \frac{z^3 + z}{(4)(df_{Error} - 2)}$$

Don’t even ask what that complex formula means! However, what it does is clear.

Ultimately, it will generate F_{Crit} for any α . You need to keep two points in mind. First, the z in the formula needs to be two-tailed, so you need to look up $.0167/2 = .0084$ in the tail of the

unit normal distribution. Second, you're generating a t value, so you need to square it to get an F_{Crit} . In this example, $z = 2.39$. So $t = 2.79$ and $F = 7.79$. Thus, using the Bonferroni procedure if you were interested in computing 6 simple pairwise comparisons on the K&W51 data set (and using $\alpha_{FW} = .10$), to be significant each $F_{Comparison}$ needs to be ≥ 7.79 .

Alternatively, you can avoid the formula entirely and use a web-based calculator, such as:

<http://www.graphpad.com/quickcalcs/Statratio1.cfm> ☺

The Sidák-Bonferroni Procedure

- This procedure is a modified Bonferroni procedure that results in a bit more power, so it is preferred to the straight Bonferroni procedure. It makes use of the following equation:

$$\alpha = 1 - (1 - \alpha_{FW})^{1/c} \quad (6.5)$$

- Because of the preference for this procedure, K&W provide useful tables in A.4. The tables illustrate information for $\alpha_{FW} = .20$, $\alpha_{FW} = .10$, $\alpha_{FW} = .05$, and $\alpha_{FW} = .01$ on pages 578-581.
- Keeping with the above example, using K&W51 and 6 comparisons with $\alpha_{FW} = .10$, we would look on p. 579. The probability associated with 6 comparisons would be .01741 (which you could also obtain by substituting .10 for α_{FW} and 6 for c in the above formula, but the table is easier). Note, of course, that if $F_{Comparison}$ yielded $p = .01741$ it would be significant with the Sidák-Bonferroni procedure, but not with the Bonferroni procedure (where it would have to be $\leq .0167$). Not a huge difference, but the Sidák-Bonferroni procedure provides a bit more power.
- You can also compare means by computing a critical mean difference according to the formula below, as applied to the above example:

$$D_{S-B} = t_{S-B} \sqrt{\frac{2MS_{Error}}{n}} = 2.76 \sqrt{\frac{2(150.458)}{4}} = 23.94$$

Thus, one possible comparison might be 4hr vs. 12hr. That difference would not be significant ($37.75 - 26.5 = 11.25$). However, in comparing 4hr vs. 20hr we would find a significant difference ($57.5 - 26.5 = 31$).

Dunnett's Test

- If you have a control group to which you wish to compare the other treatments in your study, then the Dunnett test is appropriate. Once again, the most general approach is to compute F_{Comp} and then compare that F ratio to an F_{Crit} value. For the Dunnett test, the F_{Crit} is

$$F_D = (t_D)^2$$

You look up the value of t_D in Table A.5 (pp. 582-585). To do so, you would again need to decide on the level of α_{FW} you'd like to use and then you'd need to know how many conditions are involved in your experiment (control plus experimental groups). For the K&W51 example, let's assume that the 4hr group was a control group (no sleep deprivation)

to which you'd like to compare each of the other groups. Thus, there would be a total of four groups. With $\alpha_{FW} = .10$ and 4 groups, $t_D = 2.29$. Thus, you'd compare each $F_{Comparison}$ against $F_D = 5.24$.

You could also take a critical mean difference approach with the Dunnett test:

$$D_{Dunnett} = t_{Dunnett} \sqrt{\frac{2MS_{Error}}{n}} = 2.29 \sqrt{\frac{2(150.458)}{4}} = 19.86 \quad (6.6)$$

Note that the critical mean difference here is less than that found with the Sidák-Bonferroni procedure, indicating that the Dunnett test is more powerful. Nonetheless, I would bet that you rarely find yourself in a situation where you'll want to compute the Dunnett test.

6.4 Pairwise Comparisons

Tukey's HSD Procedure

- If you are interested in comparing every pair of means (simple pairwise comparisons), you might use the Tukey HSD (Honestly Significant Difference) Procedure. Using this procedure requires you to use the Studentized Range Statistic (q) found in Appendix A.6 (pp. 586-589). Again, you can first compute $F_{Comparison}$, after which you would compare that value to a critical value obtained from the tables, which you then square. For the example we've been using (K&W51, $\alpha_{FW} = .10$, 6 pairwise comparisons):

$$F_{HSD} = \frac{q^2}{2} = \frac{3.62^2}{2} = 6.55$$

- Alternatively, you could compute a critical mean difference. For the example we've been using, you'd find:

$$D_{HSD} = q_a \sqrt{\frac{MS_{Error}}{n}} = 3.62 \sqrt{\frac{150.458}{4}} = 22.2 \quad (6.7)$$

- Note that this procedure is more liberal than the Sidák-Bonferroni procedure for what are essentially the same 6 comparisons ($D_{S-B} = 23.94$).
- K&W suggest comparing your differences among means in a matrix. For K&W51, it would look like this:

| | a ₁ (26.5) | a ₂ (37.75) | a ₃ (57.5) | a ₄ (61.75) |
|------------------------|-----------------------|------------------------|-----------------------|------------------------|
| a ₁ (26.5) | ----- | | | |
| a ₂ (37.75) | 11.25 | ----- | | |
| a ₃ (57.5) | 31.0 | 19.75 | ----- | |
| a ₄ (61.75) | 35.25 | 24.0 | 4.25 | ----- |

This table allows you to see that there are three significant comparisons.

- You cannot use this critical mean difference approach (Formula 6.7) when you have unequal sample sizes (though the formula can be modified as below) nor should you take this approach when there is heterogeneity of variance.

- When sample sizes are different, replace n in the formula with \tilde{n} , computed as:

$$\tilde{n} = \frac{2}{\frac{1}{n_1} + \frac{1}{n_2}}$$

This value is actually a special kind of mean (harmonic). Thus, if one group had $n = 10$ and the other group had $n = 20$, the \tilde{n} for Formula would be 13.33.

- If you have reason to suspect heterogeneity of variance (as discussed in Chapter 7), the formula would become:

$$D_{HSD} = q_a \sqrt{\frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right)}{2}} \quad (6.8)$$

- The df you'd use to look up q emerge from a complex formula (7.13), so we'll return to this issue once we've discussed the implications of heterogeneity of variance.

The Fisher-Hayter Procedure

- “Tukey’s procedure is the simplest way to test the pairwise differences and is the one that is most applicable to any pattern of effects.” Hmm, so why are K&W talking about alternatives? In general, some people were concerned that HSD is too conservative, so they wanted to derive more powerful simple pairwise comparison procedures.
- The Fisher-Hayter procedure requires that you first compute an overall ANOVA and reject H_0 . If you are able to reject H_0 in the overall ANOVA, then use the Studentized Range Statistic (q) found in Appendix A.6 (pp. 586-589). For HSD, you’d look up q for a treatment means. However, for Fisher-Hayter, you’d look up q for $a-1$ treatment means. Otherwise, the formulas are the same, as seen below:

$$D_{FH} = q_{a-1} \sqrt{\frac{MS_{Error}}{n}} = 3.2 \sqrt{\frac{150.458}{4}} = 19.63 \quad (6.9)$$

I’ve filled in the values that you’d use for K&W51. The critical mean difference here is smaller than that found for HSD, so this test is more powerful.

- “The Fisher-Hayter procedure provides excellent control of Type I error, a fact that has been demonstrated in several simulation studies...We suggest that you use this procedure, particularly when making calculations by hand.”

The Newman-Keuls and Related Procedures

- This procedure is sometimes referred to as Student-Newman-Keuls (SNK).
- K&W describe the procedure for computation of SNK. However, my advice (and theirs) would be to use Fisher-Hayter or Tukey’s HSD. You’ll see SNK used, but often by people who learned to use it long ago and continue to do so, even after better approaches have been identified.

6.5 Post-Hoc Error Correction

- You may be inclined to compute a whole host of comparisons, including some complex comparisons. If you're doing so in an exploratory fashion ("throw it against the wall and see what sticks"), you are asked to pay some penalty by using a very conservative test.

Scheffé's Procedure

- The Scheffé test is the most conservative post hoc test. Basically, it controls for the FW error that would occur if you were to conduct every possible comparison. Ideally, a person would be conducting far fewer comparisons than that!
- Compute the Scheffé test by first computing the F_{Comp} . In the presence of heterogeneity of variance, you would use separate variances for the denominator. In the presence of homogeneity of variance, you would use the pooled variance for the denominator. Then, test your F_{Comp} for significance by comparing it to the F_{Crit} Scheffé (F_S), where

$$F_S = (a - 1) F(df_A, df_{S/A}) \quad (6.11)$$

Thus, for comparisons in K&W51, $F_S = (3)(3.49) = 10.47$.

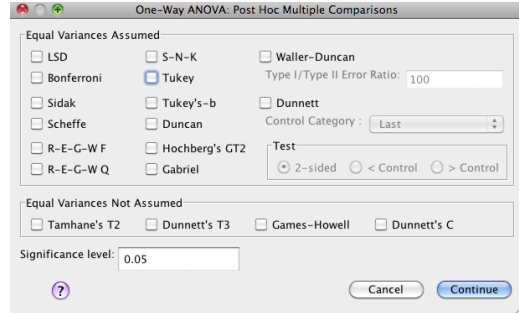
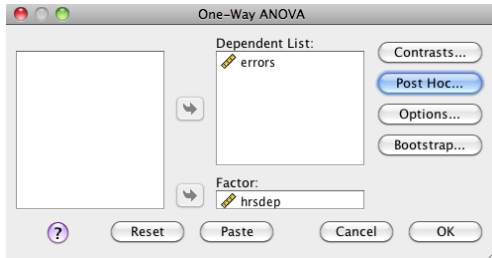
- Basically, I think that you should avoid using this procedure.

Recommendations and Guidelines (from Keppel 3rd)

- As in all things, controlling for inflated chance of familywise Type I error calls for moderation. The Sidák-Bonferroni procedure seems a reasonable approach for smaller sets of comparisons. Tukey's HSD (Tukey-Kramer) or the Fisher-Hayter procedure seem to be reasonable for simple pairwise comparisons.
- Keep in mind that FW error rate may not be as serious as it might appear to be. As Keppel notes, assuming that H_0 is true, if you replicated an experiment 2000 times and conducted the same 5 comparisons after each experiment, you would expect that a Type I error would occur in 500 experiments (.25 x 2000). However, in those 500 experiments, only 10% (50) would contain more than one Type I error in the 5 comparisons. Fewer still would have more than two.
- Furthermore, keep in mind that most experiments reflect some treatment effect (i.e., H_0 is false). That is, you are rarely dealing with a situation in which the treatment has *no* effect.
- Keppel argues for the value of planned comparisons (as does G. Loftus, 1996). Although I am in conceptual agreement, I worry about the practicalities facing a researcher who chooses to report planned comparisons. A journal editor may be sympathetic, but I worry about what an unsympathetic editor/reviewer might say about planned comparisons.
- Keppel does suggest that replications are important—especially as a means of offsetting a perceived inflated Type I error rate. Once again, in an ideal world I'd agree. However, an untenured researcher would probably benefit from doing publishable research, and journals are not yet willing to publish replications.
- We must recognize that the decision to correct for an inflated FW Type I error rate is a decision to increase the chance of making a Type II error (i.e., a decrease in power). Thus, if you have a set of planned comparisons in mind, you might well estimate your needed sample size on the basis of the planned comparisons, rather than on the basis of the overall ANOVA.

Using SPSS for Comparisons

If you choose to use Analyze->Compare Means->One-way ANOVA for analysis, you'll first see the window below left. If you click on the **Post Hoc...** button, you'll see the window below right.



As you can see, many of the procedures that K&W describe are available in SPSS. However, the Fisher-Hayter procedure isn't one of the options. Thus, if you choose to use this very reasonable post hoc procedure, you'll need to do so outside of SPSS.

Looking at the similar Tukey's HSD for the KW51 data set, you'd see the output below.

Multiple Comparisons

errors

Tukey HSD

| (I) hrsdep | (J) hrsdep | Mean Difference (I-J) | Std. Error | Sig. | 95% Confidence Interval | |
|------------|--------------|--------------------------|------------|------|-------------------------|-------------|
| | | | | | Lower Bound | Upper Bound |
| dimension2 | 1 2 | -11.250 | 8.673 | .582 | -37.00 | 14.50 |
| | dimension3 3 | -31.000* | 8.673 | .017 | -56.75 | -5.25 |
| | 4 | -35.250* | 8.673 | .007 | -61.00 | -9.50 |
| | 2 1 | 11.250 | 8.673 | .582 | -14.50 | 37.00 |
| | dimension3 3 | -19.750 | 8.673 | .158 | -45.50 | 6.00 |
| | 4 | -24.000 | 8.673 | .071 | -49.75 | 1.75 |
| dimension3 | 3 1 | 31.000* | 8.673 | .017 | 5.25 | 56.75 |
| | dimension3 2 | 19.750 | 8.673 | .158 | -6.00 | 45.50 |
| | 4 | -4.250 | 8.673 | .960 | -30.00 | 21.50 |
| | 4 1 | 35.250* | 8.673 | .007 | 9.50 | 61.00 |
| | dimension3 2 | 24.000 | 8.673 | .071 | -1.75 | 49.75 |
| | 3 | 4.250 | 8.673 | .960 | -21.50 | 30.00 |

*. The mean difference is significant at the 0.05 level.