# 1 Multiple Comparisons

Common questions:

1) After a significant ANOVA, which groups differ? (*Post hoc* tests)

2) While planning an experiment, you wish to test certain hypotheses that are a subset of the global ANOVA $H_0$. (*A priori* tests)

With the hand steadiness data, the first question arises after finding that the ANOVA global $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$ is rejected, and wishing to investigate further which sleep deprivation groups differ. The second question could arise if while planning the study, we were specifically interested, for example, in whether hand steadiness worsens as a linear function of sleep deprivation (over the range of times in the experiment). The problem that occurs in either of the two situations above is that when many individual tests are made, the chance of making a Type I error for at least one test can be far greater than the stated $\alpha$ level per test. For example, if 5 tests are made, each using $\alpha = .05$, then the probability of making a Type I error for at least one of the five tests (assuming independent tests) is $1 - (.95)^5 \approx .22$, which is far larger than .05. Multiple comparison methods aim to control a Type I error rate for a whole set of tests.

## 1.1 A guide for multiple comparisons

The author of our text does a nice job of presenting different definitions of Type I error rates that can be controlled. He presents definitions for the comparisonwise error rate, the experimentwise error rate, the false discovery rate, the strong familywise error rate, and the simultaneous confidence interval error rate. These definitions are not all standardized, however, and some texts use the term 'experimentwise error rate' to refer to what the author calls the strong familywise error rate. In deciding upon a multiple comparison method, we will take into account whether i) the contrasts to be tested are decided after collecting the data (*post hoc)* or known in advance (*a priori)*, and ii) for *a priori* contrasts whether they are orthogonal. If the contrasts to be tested are *post hoc* then we use generally more conservative methods to guard against data-snooping. For *post hoc* tests, our interest will be in methods that control one of the last two of these definitions (the strong familywise error rate and the simultaneous confidence interval error rate). For pairwise *post hoc* tests, these last two error rates are controlled by the REGWR method and Tukey's method, respectively. For non-pairwise *post hoc* tests Scheffe's method controls the strong familywise error rate and the simultaneous confidence interval error rate. In the somewhat artificial case in which a set of orthogonal contrasts has been specified *a priori*, then since the tests are independent we can simply apply separate t tests (or equivalently F tests) for each contrast, without adjusting the $\alpha$ level per contrast, in effect using the comparisonwise error rate. If a set of contrasts has been specified *a priori* but are not orthogonal, t tests are used with Bonferroni's or Holm's method. For Bonferroni's method, if $K$ tests are involved, and the overall Type I error rate is to be held at $\alpha$, then the significance level for individual tests is set at $\alpha' = \alpha/K$. Thus if 5 tests will be performed and the overall significance level for the set of tests is desired to be $\alpha = .05$, then $\alpha' = .05/5 = .01$ will be used for each individual test. This overall strategy is summarized in the following table adapted from Chapter 4 of Kirk (1995). In the table the rows identify whether the contrasts are *a priori* or *post hoc*, and the columns identify whether they are orthogonal or not. Notice that all *post hoc* contrasts are treated as if they are nonorthogonal. For the special *a priori* case of comparing all groups *versus* a control, the text discusses Dunnett's method.

|  | Orthogonal | Nonorthogonal |
|---|---|---|
| *A priori* | Separate t-tests | Separate t-tests with Bonferroni's or Holm's method |
| *Post hoc* |  | Pairwise: REGWR or Tukey; Non-pairwise: Scheffe |

# 2 Methods

## 2.1 Holm's method

Like Bonferroni's method, except that not all tests are compared to $\alpha/K$. The p values of the tests are ordered from smallest to largest, and then the smallest is compared to $\alpha/K$. If significant, then the next smallest p value is compared to $\alpha/(K-1)$. This pattern is repeated as long as the p values are all significant. As soon as one is not significant, the process stops and all remaining tests are nonsignificant.

## 2.2 Tukey's method for pairwise contrasts

Tukey's method is used for testing all pairwise differences between groups. Here we assume that the sample size is equal in each group, which is represented by $n$. Modifications are available when sample sizes differ between groups, as shown in the text. To perform Tukey's method, follow these steps:

1. Rank the sample means.

2. Calculate $HSD = q_\alpha(g,\nu)\sqrt{\frac{MS_E}{n}}$, where $q_\alpha(g,\nu)$ is the $100(\alpha)\%$ point of the Studentized range distribution, $g$ is the number of groups, $\nu$ is the degrees of freedom for $MS_E$, and $n$ is the common sample size per group.

3. Two population means $\mu_i$ and $\mu_{i'}$ are declared different if $|\overline{y}_{i\cdot} - \overline{y}_{i'\cdot}| \geq HSD$.

4. The results for the set of groups are often depicted graphically by drawing the ordered means and connecting groups that do not differ by a line.

Confidence intervals for $\mu_i$ - $\mu_{i'}$ can be constructed as $(\overline{y}_{i\cdot} - \overline{y}_{i'\cdot}) \pm HSD$. The Studentized range distribution used by Tukey is the distribution of the difference $\overline{y}_{MAX} - \overline{y}_{MIN}$ for a given number $(g)$ of groups. In effect it is treating all pairwise comparisons as if they came from data snooping to pick the most different pair of means. Thus it controls the strong familywise Type I error rate at $\alpha$ for the entire collection of pairwise tests. It cannot disagree with the result of the global ANOVA test, in the sense that if any pair of means are declared different by Tukey, then the global ANOVA $H_0$ must have been rejected.

## 2.3 The REGWR method

Similar to Tukey's method, except that it is a step-down method, meaning that it uses different studentized range values depending upon how close the sample means are for the groups being compared.

## 2.4 Scheffe's method for general contrasts

For a general contrast $L = \sum_{i=1}^{g} w_i\mu_i$ Scheffe's method can be used to either test a hypothesis about $L$ or construct a confidence interval. To test $H_0 : L = 0$, against $H_A : L \neq 0$ with Scheffe's method, follow these steps:

1. Calculate $\widehat{L} = \sum_{i=1}^{g} w_i\overline{y}_{i\cdot}$ .

2. Calculate $S = \sqrt{(g-1)F_{\alpha,g-1,\nu}}\sqrt{\widehat{Var}(\widehat{L})}$, where $\nu$ is the degrees of freedom for $MS_E$, and $\widehat{Var}(\widehat{L})$ is from the notes on contrasts.

3. If $|\widehat{L}| > S$ then reject $H_0$.

Confidence intervals for $L$ can be constructed as $\widehat{L} \pm S$. Scheffe's method controls the (strong familywise) Type I error rate at $\alpha$ for the entire collection of general contrasts, whether pairwise or nonpairwise. Like Tukey's method it cannot disagree with the result of the global ANOVA $H_0$. Scheffe's method is too conservative for use with pairwise comparisons.

## 3 Reference

Kirk, R. E. (1995) Experimental Design: Procedures for the Behavioral Sciences. Pacific Grove: Brooks/Cole.