1 The need to control error rates with multiple comparisons

As shown in the text, when a set of \( n \) hypotheses are tested where each has a Type I error rate of \( \alpha_C \) (the comparisonwise error rate), then an upper limit on the probability that at least one of the tests commits a Type I error (called the experimentwise error rate) is:

\[
\alpha_E = 1 - (1 - \alpha_C)^n
\]

Some values for these error rates are compared in Table 3.8 of the text, and they show that \( \alpha_E \) increases rapidly as more tests are conducted.

2 A simple guide for multiple comparisons

There are a vast number of methods used for multiple comparison tests, and we will only consider a small number of them. We will only consider in detail three types of multiple comparison tests: t tests, Tukey’s method, and Scheffe’s method. Fisher’s LSD method will also be discussed since it is widely used. The choice between these methods is governed by the type of contrasts being tested. In the somewhat artificial case in which a set of orthogonal contrasts has been specified \textit{a priori}, then since the tests are independent we can simply apply separate t tests for each contrast, without adjusting the \( \alpha \) level per contrast (this recommendation is at odds with the author of our text). If a set of contrasts has been specified \textit{a priori} but are not orthogonal, t tests are again used but with a Bonferroni correction. In this case if \( n \) tests are involved, and the experimentwise error rate is to be held at \( \alpha \), then the comparisonwise error rate (for individual tests) is set at \( \alpha' = \alpha/n \). 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. If the contrasts to be tested are decided after collecting the data (post hoc) then we use generally more conservative methods to guard against data-snooping. For pairwise contrasts we can use Tukey’s method and for non-pairwise contrasts we use Scheffe’s method. This overall strategy is summarized in the following table, where the rows identify whether the contrasts are \textit{a priori} or \textit{post hoc}, and the columns identify whether they are orthogonal or not. Notice that all \textit{post hoc} contrasts are treated as if they are nonorthogonal. The text considers some other alternative multiple comparison methods. If one treatment is considered a control, so that the pairwise contrasts of interest involve differences with the control, then Dunnett’s method can be used. A related idea is to compare all treatments to the best (largest or smallest, depending on context) treatment, this is called the \textbf{multiple comparisons with the best} procedure. Some methods use multiple criteria to declare differences between treatment means based on the ordering of the sample means, these are called multiple range tests and the text presents the \textbf{Student-Newman-Keuls multiple range test}.

<table>
<thead>
<tr>
<th>Orthogonal</th>
<th>Nonorthogonal</th>
</tr>
</thead>
<tbody>
<tr>
<td>\textit{A priori}</td>
<td>Separate t-tests</td>
</tr>
<tr>
<td>\textit{Post hoc}</td>
<td>Pairwise: Tukey; Non-pairwise: Scheffe</td>
</tr>
</tbody>
</table>

3 Methods

3.1 \textit{t} tests

We can estimate the contrast
\[ C = \sum_{i=1}^{t} k_i \mu_i \text{ with } c = \sum_{i=1}^{t} k_i \bar{y}_i \text{ and } \widehat{\text{Var}}(c) = \text{MSE} \sum_{i=1}^{t} k_i^2 r_i, \]

which leads to a t test of \( H_0 : C = \sum_{i=1}^{t} k_i \mu_i = 0, \)

\[ t = \frac{c}{s_c} = \frac{\sum_{i=1}^{t} k_i \bar{y}_i}{\sqrt{\text{Var}(c)}}. \]

For a two-tailed test, the t value is compared to \( t_{n-1, v}, \) where \( v \) is the degrees of freedom for \( \text{MSE} \) (\( df = t(n-1) \) for balanced 1 way ANOVA). Confidence intervals for \( C \) can also be constructed as \( c \pm t_{\alpha/2, v} s_c. \)

### 3.2 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 \( r \). Modifications are available when sample sizes differ between groups. To perform Tukey’s method, follow these steps:

1. Rank the sample means.
2. Calculate \( HSD = q_{\alpha, t, v} \sqrt{\frac{\text{MSE}}{r}} \), where \( q_{\alpha, t, v} \) is the Studentized range statistic, \( t \) is the number of groups, \( v \) is the degrees of freedom for \( \text{MSE} \), and \( r \) is the common sample size per group.
3. Two population means \( \mu_i \) and \( \mu_j \) are declared different if \( |\bar{y}_i - \bar{y}_j| \geq HSD. \)
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 \( \mu_i - \mu_j \) can be constructed as \( (\bar{y}_i - \bar{y}_j) \pm HSD. \) The Studentized range distribution used by Tukey is the distribution of the difference \( \bar{y}_{\text{MAX}} - \bar{y}_{\text{MIN}} \) for a given number \( (t) \) 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 Type I error rate at \( \alpha \) for the entire collection of pairwise tests.

### 3.3 Scheffe’s method for general contrasts

For a general contrast \( C = \sum_{i=1}^{t} k_i \mu_i \) Scheffe’s method can be used to either test a hypothesis about \( C \) or construct a confidence interval. To test \( H_0 : C = 0, \) against \( H_A : C \neq 0 \) with Scheffe’s method, follow these steps:

1. Calculate \( c = \sum_{i=1}^{t} k_i \bar{y}_i. \)
2. Calculate \( S = s_c \sqrt{(t-1)F_{\alpha, t-1, v}}, \) where \( v \) is the degrees of freedom for \( \text{MSE}. \)
3. If \( |c| > S \) then reject \( H_0. \)

Confidence intervals for \( C \) can be constructed as \( c \pm S. \) Scheffe’s method controls the 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 considered to be too conservative for use with pairwise comparisons.

### 3.4 Fisher’s LSD method for pairwise contrasts

Another popular method for pairwise comparisons is Fisher’s LSD (Least significant difference) method. In this method the means are ranked and then differences \( (\bar{y}_i - \bar{y}_j) \) are compared to \( \text{LSD} = t_{\alpha/2, v} \sqrt{\text{MSE} \left( \frac{1}{r_i} + \frac{1}{r_j} \right)}. \)
Many researchers advocate only performing Fisher’s LSD method after rejecting the global ANOVA $H_0$, in which case it is described as \textbf{Fisher’s protected LSD method}. Notice that if a pairwise contrast is being used, then Scheffe’s $S$ value reduces to $S = \sqrt{(t-1)F_{\alpha,t-1,u}}\sqrt{MSE\left(\frac{1}{r_i} + \frac{1}{r_j}\right)}$. A comparison of the $LSD$ and $S$ terms shows that Fisher’s method uses $t_{\alpha/2,u} = \sqrt{t_{\alpha/2,u}^2} = \sqrt{F_{\alpha,1,u}}$, which differs from Scheffe’s method in two ways: by not having a $(t - 1)$ multiplier next to the $F$ value, and by having only 1 numerator degree of freedom instead of $t - 1$. Fisher’s LSD method is therefore much more liberal than Scheffe’s method, and in fact does not always control the Type I error rate at $\alpha$ even when only the protected method is used.

4 \hspace{1cm} \textbf{A final note}

As previously stated, there are a vast number of multiple comparison methods in use. We have only discussed a very small number of methods that are widely recognized, implemented in most software, and all provide confidence intervals as well as hypothesis tests. There are, for example, methods for pairwise contrasts that control Type I error for the collection of tests as Tukey’s method does, but have much greater power for detecting differences. A good discussion of many methods is found in Chapter 4 of Kirk (1995). The methods discussed above were first developed for studies where a small or moderate number of comparisons are made. In some current scientific studies, much larger numbers of tests are made, such as in genomic studies. For these situations the concept of experimentwise error is usually no longer useful, and other criteria such as the false discovery rate (the expected proportion of false positives among the positive findings) are used instead.

5 \hspace{1cm} \textbf{Reference}