1 Chapter 3: More on Treatment Comparisons

1.1 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 α_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 α_E increases rapidly as more tests are conducted.

1.2 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 *a priori*, then since the tests are independent we can simply apply separate t tests for each contrast, without adjusting the α level per contrast (this recommendation is at odds with the author of our text). If a set of contrasts has been specified *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 α, 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 *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. 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 **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 **Student-Newman-Keuls multiple range test**.

	Orthogonal	Nonorthogonal
A priori	Separate t-tests	Separate t-tests with Bonferroni correction
Post hoc		Pairwise: Tukey; Non-pairwise: Scheffe

1.3 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.

2 Chapter 4: Model assumptions and transformations

2.1 Valid results depend on model assumptions

This point is discussed at the beginnnig of Chapter 4. Some general results on effects of departures from assumptions are mentioned, particularly problems that occur when variances are unequal and sample sizes are also unequal.

2.2 Model diagnostics: residual analysis

Recall that the residuals e_{ij} in the ANOVA model $y_{ij} = \mu_i + e_{ij}$ should 1) have a mean of 0, 2) be independent, 3) have a normal distribution, and 4) have a common variance σ^2. We can estimate the e_{ij} terms from our sample via one of several types of residuals:

Name of residual	**Formula**	**SAS Proc GLM keyword**	**R function**
residual	$\widehat{e}_{ij} = y_{ij} - \widehat{\mu}_{i.} = y_{ij} - \overline{y}_{i.}$	r = or residual =	residuals()
standardized residual	$w_{ij} = \widehat{e}_{ij}/\sqrt{MSE}$		
studentized residual	$\widetilde{e}_{ij} = \widehat{e}_{ij}/\sqrt{MSE(1-1/r_i)}$	student =	rstandard()
jackknife residual	$\widetilde{e}_{(-ij)} = \widehat{e}_{ij}/\sqrt{MSE_{(-ij)}(1-1/r_i)}$	rstudent =	rstudent()

where $MSE_{(-ij)}$ is MSE computed without the jth observation in the ith group. The last three types (standardized, studentized, and jackknife) are fairly similar if the model assumptions are satisfied. Since the last type, the jackknife residual, is best at detecting various problems, we will often use it. It can be obtained in SAS Proc GLM on the Output statement with the RSTUDENT option. A sound strategy for diagnosing the adequacy of a linear model is to obtain a variety of plots of residuals. These include histograms, normal plots, and scatter plots of the residuals versus the predicted values. A useful way to evaluate the collection of residual plots is to look at the plots for problems, and focus on characterizing the potential effect of these problems on the model. In addition to the various types of residuals shown above, the text discusses the use of the square root of the absolute residuals in a plot with the predicted values, and also some tests for homogeneity of variance, such as the Brown-Forsythe modification of the Levene test.

2.3 Transformations to help satisfy model assumptions

The text discusses different approaches for obtaining transformations of data to help meet model assumptions, focusing primarily on two methods, i) variance stabilizing transformations for known distributions, and ii) use of the Box-Cox power transformation.

3 Chapter 5: Fixed versus Random effects models

So far all of the treatment effects that we have studied were fixed, meaning that the observed levels of the treatment factor are the only levels of interest. In this model,

$$y_{ij} = \mu + \tau_i + e_{ij},$$

the t values of τ_i, $\tau_1, \tau_2, ..., \tau_t$, are the only effects of interest, and e_{ij} is the only random term, e_{ij} has a normal distribution with mean 0 and variance σ_e^2 ($e_{ij} \sim N(0, \sigma_e^2)$). However, sometimes the treatment levels under study are only a sample of the possible treatment levels, and we wish to generalize our results to this larger set of levels. The high temperature alloy casting example is like this, where the three fabrication castings in the study are only a sample of the set of fabrication castings. For this type of study a more appropriate model is the random effects model:

$$y_{ij} = \mu + a_i + e_{ij},$$

Here the a_i terms are now random, $a_i \sim N(0\ ,\ \sigma_a^2\ )$, the e_{ij} terms have the same assumptions as before, and the a_i and e_{ij} are independent. Now the a_i 's are viewed as a random sample from a population of a_i 's, and the ANOVA H_0 is H_0: $\sigma_a^2 = 0$ vs. $H_a : \sigma_a^2 > 0$.

How do you know if an effect is fixed or random?

1. How were the levels for the factor in question chosen?
2. Is it desired to generalize the results to levels that weren't used?

Implications for random effects

1. Generally you are not interested in doing multiple comparisons. There may be interest in estimating σ_a^2 , and σ_e^2 , or functions of them, like ratios.
2. The denominator of the F statistic may change (this occurs in more complex designs).

Analysis of random effects models

The null hypothesis, H_0: $\sigma_a^2 = 0$ vs. $H_a : \sigma_a^2 > 0$ can be tested by the ANOVA F statistic just as with the fixed effects model, as can be seen from an examination of the expected mean squares. The text obtains estimators of the variance components by equating observed mean squares to their expected values, then solving for the variance components. This method, often called the method of moments, can encounter problems such as negative variance component estimates. Instead, we will use REML (restricted maximum likelihood, similar to maximum likelihood) estimators that are computed using Proc MIXED in SAS.

4 Reference

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