Chapter 5
Multiple Comparison Techniques

In analyzing a linear model we can examine as many single degree of freedom hypotheses as we want. If we test all of these hypotheses at, say, the 0.05 level, then the (weak) *experimentwise error rate* (the probability of rejecting at least one of these hypotheses when all are true) will be greater than 0.05. Multiple comparison techniques are methods of performing the tests so that if all the hypotheses are true, then the probability of rejecting any of the hypotheses is no greater than some specified value, i.e., the experimentwise error rate is controlled.

A multiple comparison method can be said to be more powerful than a competing method if both methods control the experimentwise error rate at the same level, but the method in question rejects hypotheses more often than its competitor. Being more powerful, in this sense, is a mixed blessing. If one admits the idea that a null hypothesis really can be true (an idea that I am often loath to admit), then the purpose of a multiple comparison procedure is to identify which hypotheses are true and which are false. The more powerful of two multiple comparison procedures will be more likely to correctly identify hypotheses that are false as being false. It will also be more likely to incorrectly identify hypotheses that are true as being false.

A related issue is that of examining the data before deciding on the hypotheses. If the data have been examined, an hypothesis may be chosen to test because it looks as if it is likely to be significant. The nominal significance level of such a test is invalid. In fact, when doing multiple tests by any standard method, nearly all the nominal significance levels are invalid. For some methods, however, selecting the hypotheses after examining the data make the error levels intolerably bad.

The sections of this chapter contain discussions of individual multiple comparison methods. The methods discussed are Scheffé's method, the Least Significant Difference (LSD) method, the Bonferroni method, Tukey's Honest Significant Difference (HSD) method, and multiple range tests. The section on multiple range tests examines both the Newman–Keuls method and Duncan's method. The final section of the chapter compares the various methods. For a more complete discussion of multiple comparison methods, see Miller (1981) or, more recently, Hochberg and Tamhane (1987) or Hsu (1996). Miller's book includes several methods that are not discussed here. Christensen (1996a) discusses the methods of this chapter at a more

R. Christensen, *Plane Answers to Complex Questions: The Theory of Linear Models*, Springer Texts in Statistics, DOI 10.1007/978-1-4419-9816-3_5,

applied level, discusses Dunnett's method for comparing treatments with a control, and discusses Ott's analysis of means method.

5.1 Scheffé's Method

Scheffé's method of multiple comparisons is an omnibus technique that allows one to test any and all single degree of freedom hypotheses that put constraints on a given subspace. It provides the assurance that the experimentwise error rate will not exceed a given level α. Typically, this subspace will be for fitting a set of parameters after fitting the mean, and in ANOVA problems is some sort of treatment space.

It is, of course, easy to find silly methods of doing multiple comparisons. One could, for example, always accept the null hypothesis. However, if the subspace is of value in fitting the model, Scheffé's method assures us that there is at least one hypothesis in the subspace that will be rejected. That is, if the F test is significant for testing that the subspace adds to the model, then there exists a linear hypothesis, putting a constraint on the subspace, that will be deemed significant by Scheffé's method.

Scheffé's method is that an hypothesis $H_0 : \lambda'\beta = 0$ is rejected if

$$\frac{SS(\lambda'\beta)/s}{MSE} > F(1-\alpha, s, dfE),$$

where $SS(\lambda'\beta)$ is the sum of squares for the usual test of the hypothesis, s is the dimension of the subspace, and $\lambda'\beta = 0$ is assumed to put a constraint on the subspace.

In terms of a one-way analysis of variance where the subspace is the space for testing equality of the treatment means, Scheffé's method applies to testing whether all contrasts equal zero. With t treatments, a contrast is deemed significantly different from zero at the α level if the sum of squares for the contrast divided by $t-1$ and the MSE is greater than $F(1-\alpha, t-1, dfE)$.

Theorem 5.1.1 given below leads immediately to the key properties of Scheffé's method. Recall that if $\rho'X\beta = 0$ puts a constraint on a subspace, then $M\rho$ is an element of that subspace. Theorem 5.1.1 shows that the F test for the subspace rejects if and only if the Scheffé test rejects the single degree of freedom hypothesis $\rho'X\beta = 0$ for some ρ with $M\rho$ in the subspace. The proof is accomplished by finding a vector in the subspace having the property that the sum of squares for testing the corresponding one degree of freedom hypothesis equals the sum of squares for the entire space. Of course, the particular vector that has this property depends on Y. To emphasize this dependence on Y, the vector is denoted m_Y. In the proof of the theorem, m_Y is seen to be just the projection of Y onto the subspace (hence the use of the letter m in the notation).

It follows that for a one-way ANOVA there is always a contrast for which the contrast sum of squares equals the sum of squares for treatments. The exact nature of this contrast depends on Y and often the contrast is completely uninterpretable.

Nevertheless, the existence of such a contrast establishes that Scheffé's method rejects for some contrast if and only if the test for equality of treatments is rejected.

Theorem 5.1.1. Consider the linear model $Y = X\beta + e$ and let M_* be the perpendicular projection operator onto some subspace of $C(X)$. Let $r(M_*) = s$. Then

$$\frac{Y'M_*Y/s}{MSE} > F(1-\alpha, s, dfE)$$

if and only if there exists a vector m_Y such that $Mm_Y \in C(M_*)$ and

$$\frac{SS(m_Y'X\beta)/s}{MSE} > F(1-\alpha, s, dfE).$$

PROOF. $\Rightarrow$ We want to find a vector m_Y so that if the F test for the subspace is rejected, then Mm_Y is in $C(M_*)$, and the hypothesis $m_Y'X\beta = 0$ is rejected by Scheffé's method. If we find m_Y within $C(M_*)$ and $SS(m_Y'X\beta) = Y'M_*Y$, we are done. Let $m_Y = M_*Y$.

As in Section 3.5, $SS(m_Y'X\beta) = Y'M_*m_Y[m_Y'M_*m_Y]^{-1}m_Y'M_*Y$. Since $M_*m_Y = M_*M_*Y = M_*Y = m_Y$, we have $SS(m_Y'X\beta) = Y'm_Y[m_Y'm_Y]^{-1}m_Y'Y = (Y'M_*Y)^2/Y'M_*Y = Y'M_*Y$, and we are finished.

$\Leftarrow$ We prove the contrapositive, i.e., if $Y'M_*Y/s\,MSE \leq F(1-\alpha, s, dfE)$, then for any ρ such that $M\rho \in C(M_*)$, we have $SS(\rho'X\beta)/s\,MSE \leq F(1-\alpha, s, dfE)$. To see this, observe that

$$SS(\rho'X\beta) = Y'[M\rho(\rho'M\rho)^{-1}\rho'M]Y.$$

Since $[M\rho(\rho'M\rho)^{-1}\rho'M]$ is the perpendicular projection matrix onto a subspace of $C(M_*)$,

$$Y'[M\rho(\rho'M\rho)^{-1}\rho'M]Y \leq Y'M_*Y$$

and we are done. □

$Y'M_*Y$ is the sum of squares for testing the reduced model $Y = (M - M_*)\gamma + e$. If this null model is true,

$$\Pr\left[\frac{Y'M_*Y}{s\,MSE} > F(1-\alpha, s, dfE)\right] = \alpha.$$

The theorem therefore implies that the experimentwise error rate for testing all hypotheses $\rho'X\beta = 0$ with $M\rho \in C(M_*)$ is exactly α. More technically, we wish to test the hypotheses

$$H_0 : \lambda'\beta = 0 \quad \text{for } \lambda \in \{\lambda | \lambda' = \rho'X \text{ with } M\rho \in C(M_*)\}.$$

The theorem implies that

$$\Pr\left[\frac{SS(\lambda'\beta)/s}{MSE} > F(1-\alpha,s,dfE) \text{ for some } \lambda, \lambda' = \rho' X, M\rho \in C(M_*)\right] = \alpha,$$

so the experimentwise error rate is α. The theorem also implies that if the omnibus F test rejects, there exists some single degree of freedom test that will be rejected. Note that which single degree of freedom tests are rejected depends on what the data are, as should be expected.

Scheffé's method can also be used for testing a subset of the set of all hypotheses putting a constraint on $C(M_*)$. For testing a subset, the experimentwise error rate will be no greater than α and typically much below α. The primary problem with using Scheffé's method is that, for testing a finite number of hypotheses, the experimentwise error rate is so much below the nominal rate of α that the procedure has very little power. (On the other hand, you can be extra confident, when rejecting with Scheffé's method, that you are not making a type I error.)

Suppose that we want to test

$$H_0 : \lambda_k'\beta = 0, \quad k = 1, \ldots, r.$$

The constraints imposed by these hypotheses are $M\rho_k$, $k = 1, \ldots, r$, where $\lambda_k' = \rho_k' X$. If $C(M_*)$ is chosen so that $C(M\rho_1, \ldots, M\rho_r) \subset C(M_*)$, then by the previous paragraph, if H_0 is true,

$$\Pr\left[\frac{SS(\lambda_k'\beta)/s}{MSE} > F(1-\alpha,s,dfE) \text{ for some } k,\ k = 1, \ldots, r\right] \leq \alpha.$$

For testing a finite number of hypotheses, it is possible to reject the overall F test but not reject for any of the specific hypotheses.

We now show that the most efficient procedure is to choose $C(M\rho_1, \ldots, M\rho_r) = C(M_*)$. In particular, given that a subspace contains the necessary constraints, the smaller the rank of the subspace, the more powerful is Scheffé's procedure. Consider two subspaces, one of rank s and another of rank t, where $s > t$. Both procedures guarantee that the experimentwise error rate is no greater than α. The more powerful procedure is the one that rejects more often. Based on the rank s subspace, Scheffé's method rejects if

$$SS(\lambda'\beta)/MSE > sF(1-\alpha,s,dfE).$$

For the rank t subspace, the method rejects if

$$SS(\lambda'\beta)/MSE > tF(1-\alpha,t,dfE).$$

With $s > t$, by Theorem C.4,

$$sF(1-\alpha,s,dfE) \geq tF(1-\alpha,t,dfE).$$

One gets more rejections with the rank t space, hence it gives a more powerful procedure.

EXAMPLE 5.1.2. *One-Way ANOVA.*
Consider the model

$$y_{ij} = \mu + \alpha_i + e_{ij}, \quad e_{ij}\text{s i.i.d. } N(0,\sigma^2),$$

$i = 1,2,3,4$, $j = 1,\ldots,N$. To test the three contrast hypotheses

$$\begin{aligned}\lambda_1'\beta &= \alpha_1 + \alpha_2 - \alpha_3 - \alpha_4 = 0,\\ \lambda_2'\beta &= \alpha_1 - \alpha_2 + \alpha_3 - \alpha_4 = 0,\\ \lambda_3'\beta &= \alpha_1 + 0 + 0 - \alpha_4 = 0,\end{aligned}$$

we can observe that the contrasts put constraints on the space for testing $H_0 : \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4$ and the space has rank 3. We can apply Scheffé's method: reject $H_0 : \lambda_k'\beta = 0$ if

$$\frac{SS(\lambda_k'\beta)/3}{MSE} > F(1-\alpha,3,4(N-1)).$$

A more efficient method is to notice that $\lambda_1'\beta + \lambda_2'\beta = 2\lambda_3'\beta$. This is true for any β, so $\lambda_1' + \lambda_2' = 2\lambda_3'$ and, using (4.2.2), $M\rho_1 + M\rho_2 = 2M\rho_3$. Since λ_1 and λ_2 are linearly independent, $M\rho_1$ and $M\rho_2$ are also; thus $C(M\rho_1, M\rho_2, M\rho_3)$ is a rank 2 space and Scheffé's method can be applied as: reject $H_0 : \lambda_k'\beta = 0$ if

$$\frac{SS(\lambda_k'\beta)/2}{MSE} > F(1-\alpha,2,4(N-1)).$$

One virtue of Scheffé's method is that since it is really a test of all the hypotheses in a subspace, you can look at the data to help you pick an hypothesis and the test remains valid.

Scheffé's method can also be used to find simultaneous confidence intervals. To show this we need some additional structure for the problem. Let $X = [X_0, X_1]$ and let $\beta' = [\beta_0', \beta_1']$, so that

$$Y = X_0\beta_0 + X_1\beta_1 + e.$$

Let M and M_0 be the perpendicular projection operators onto $C(X)$ and $C(X_0)$, respectively, and let $M_* = M - M_0$. We seek to find simultaneous confidence intervals for all estimable functions $\rho' X_1\beta_1$. Note that $\rho' X_1\beta_1$ is estimable if and only if $\rho' X_0 = 0$, which occurs if and only if $0 = M_0\rho = M\rho - M_*\rho$, i.e., $M\rho = M_*\rho$. It follows that if $\rho' X_1\beta_1$ is an estimable function, then $\rho' X_1\beta_1 = \rho' MX_1\beta_1 = \rho' M_* X_1\beta_1$. Conversely, for any vector ρ, $\rho' M_*(X_0\beta_0 + X_1\beta_1) = \rho' M_* X_1\beta_1$ is an estimable function. Proceeding as in Section 3.7, and observing that $M_* X\beta = M_* X_0\beta_0 + M_* X_1\beta_1 = M_* X_1\beta_1$, we have

$$\frac{(Y - X_1\beta_1)' M_*(Y - X_1\beta_1)/r(M_*)}{MSE} \sim F(r(M_*), dfE, 0),$$

so that

$$\Pr\Big[\frac{(Y-X_1\beta_1)'M_*(Y-X_1\beta_1)/r(M_*)}{MSE} \le F(1-\alpha, r(M_*), dfE)\Big] = 1-\alpha$$

or, equivalently,

$$1-\alpha = \Pr\Bigg[\frac{(Y-X_1\beta_1)'M_*\rho(\rho'M_*\rho)^{-1}\rho'M_*(Y-X_1\beta_1)/r(M_*)}{MSE} \le F(1-\alpha, r(M_*), dfE) \text{ for all } \rho\Bigg]$$

$$= \Pr\big[|\rho'M_*Y - \rho'M_*X_1\beta_1| \le \sqrt{(\rho'M_*\rho)(MSE)r(M_*)F(1-\alpha, r(M_*), dfE)} \text{ for all } \rho\big].$$

This leads to obvious confidence intervals for all functions $\rho'M_*X_1\beta_1$ and thus to confidence intervals for arbitrary estimable functions $\rho'X_1\beta_1$.

EXAMPLE 5.1.3. *One-Way ANOVA*.
Consider the model $y_{ij} = \mu + \alpha_i + e_{ij}$, e_{ij}s independent $N(0,\sigma^2)$, $i = 1,\ldots,t$, $j = 1,\ldots,N_i$, and the space for testing $\alpha_1 = \alpha_2 = \cdots = \alpha_t$. The linear functions that put constraints on that space are the contrasts. Scheffé's method indicates that $H_0 : \sum_{i=1}^t \lambda_i\alpha_i = 0$ should be rejected if

$$\frac{(\sum \lambda_i\bar{y}_{i\cdot})^2/(\sum \lambda_i^2/N_i)}{(t-1)MSE} > F(1-\alpha, t-1, dfE).$$

To find confidence intervals for contrasts, write $X = [J, X_1]$ and $\beta' = [\mu, \beta_1']$, where $\beta_1' = [\alpha_1,\ldots,\alpha_t]$. We can get simultaneous confidence intervals for estimable functions $\rho'X_1\beta_1$. As discussed in Chapter 4, the estimable functions $\rho'X_1\beta_1$ are precisely the contrasts. The simultaneous $(1-\alpha)100\%$ confidence intervals have limits

$$\sum \lambda_i\bar{y}_{i\cdot} \pm \sqrt{(t-1)F(1-\alpha, t-1, dfE)MSE\left(\sum \lambda_i^2/N_i\right)}.$$

5.2 Least Significant Difference Method

The Least Significant Difference (LSD) method is a general technique for testing a fixed number of hypotheses $\lambda_k'\beta = 0$, $k = 1,\ldots,r$, chosen without looking at the data. The constraints imposed by these hypotheses generate some subspace. (Commonly, one identifies the subspace first and picks hypotheses that will generate it.) The technique is a simple two-stage procedure. First, do an α level F test for whether the subspace adds to the model. If this omnibus F test is not significant, we can conclude that the data are consistent with $\lambda_k'\beta = 0$, $k = 1,\ldots,r$. If the F test is significant, we want to identify which hypotheses are not true. To do this, test each hypothesis $\lambda_k'\beta = 0$ with a t test (or an equivalent F test) at the α level.

The experimentwise error rate is controlled by using the F test for the subspace. When all of the hypotheses are true, the probability of identifying any of them as false is no more than α, because α is the probability of rejecting the omnibus F test. Although the omnibus F test is precisely a test of $\lambda_k'\beta = 0$, $k = 1, \ldots, r$, even if the F test is rejected, the LSD method may not reject any of the specific hypotheses being considered. For this reason, the experimentwise error rate is less than α.

The LSD method is more powerful than Scheffé's method. If the hypotheses generate a space of rank s, then Scheffé's method rejects if $SS(\lambda_k'\beta)/MSE > sF(1-\alpha, s, dfE)$. The LSD rejects if $SS(\lambda_k'\beta)/MSE > F(1-\alpha, 1, dfE)$. By Theorem C.4, $sF(1-\alpha, s, dfE) > F(1-\alpha, 1, dfE)$, so the LSD method will reject more often than Scheffé's method. Generally, the LSD method is more powerful than other methods for detecting when $\lambda_k'\beta \neq 0$; but if $\lambda_k'\beta = 0$, it is more likely than other methods to incorrectly identify the hypothesis as being different from zero.

Note that it is not appropriate to use an F test for a space that is larger than the space generated by the r hypotheses. Such an F test can be significant for reasons completely unrelated to the hypotheses, thus invalidating the experimentwise error rate.

Exercise 5.1 Consider the ANOVA model

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

$i = 1, \ldots, t$, $j = 1, \ldots, N$, with the e_{ij}s independent $N(0, \sigma^2)$. Suppose it is desired to test the hypotheses $\alpha_i = \alpha_{i'}$ for all $i \neq i'$. Show that there is one number, called the LSD, so that the least significant difference rejects $\alpha_i = \alpha_{i'}$ precisely when

$$|\bar{y}_{i\cdot} - \bar{y}_{i'\cdot}| > LSD.$$

Exercise 5.2 In the model of Exercise 5.1, let $t = 4$. Suppose we want to use the LSD method to test contrasts defined by

Name	λ_1	λ_2	λ_3	λ_4
A	1	1	-1	-1
B	0	0	1	-1
C	$1/3$	$1/3$	$1/3$	-1

Describe the procedure. Give test statistics for each test that is to be performed.

5.3 Bonferroni Method

Suppose we have chosen, before looking at the data, a set of r hypotheses to test, say, $\lambda_k'\beta = 0$, $k = 1, \ldots, r$. The Bonferroni method consists of rejecting $H_0 : \lambda_k'\beta = 0$ if

$$\frac{SS(\lambda_k'\beta)}{MSE} > F\left(1-\frac{\alpha}{r},1,dfE\right).$$

The Bonferroni method simply reduces the significance level of each individual test so that the sum of the significance levels is no greater than α. (In fact, the reduced significance levels do not have to be α/r as long as the sum of the individual significance levels is α.)

This method rests on a Bonferroni inequality. For sets $A_1,\ldots,A_r$, $\Pr(\bigcup_{k=1}^r A_k) \le \sum_{k=1}^r \Pr(A_k)$. (This inequality is nothing more than the statement that a probability measure is finitely subadditive.) If all the hypotheses $\lambda_k'\beta = 0$, $k = 1,\ldots,r$ are true, then the experimentwise error rate is

$$\begin{aligned}
&\Pr\left(SS(\lambda_k'\beta) > MSE\ F\left(1-\frac{\alpha}{r},1,dfE\right) \text{ for some } k\right)\\
&= \Pr\left(\bigcup_{k=1}^{r}\left[SS(\lambda_k'\beta) > MSE\ F\left(1-\frac{\alpha}{r},1,dfE\right)\right]\right)\\
&\le \sum_{k=1}^{r}\Pr\left(SS(\lambda_k'\beta) > MSE\ F\left(1-\frac{\alpha}{r},1,dfE\right)\right)\\
&= \sum_{k=1}^{r}\frac{\alpha}{r} = \alpha.
\end{aligned}$$

If the hypotheses to be tested are chosen after looking at the data, the individual significance levels of α/r are invalid, so the experimentwise error rate has not been controlled.

Given that the subspace F test is rejected, the LSD method is more powerful than the Bonferroni method because $F\left(1-\frac{\alpha}{r},1,dfE\right) > F(1-\alpha,1,dfE)$. The Bonferroni method is designed to handle a finite number of hypotheses, so it is not surprising that it is usually a more powerful method than Scheffé's method for testing the r hypotheses if r is not too large.

5.4 Tukey's Method

Tukey's method, also known as the Honest Significant Difference (HSD) method, is designed to compare all pairs of means for a set of independent normally distributed random variables with a common variance. Let $y_i \sim N(\mu_i,\sigma^2)$, $i = 1,\ldots,t$, let the y_is be independent, and let S^2 be an estimate of σ^2 with S^2 independent of the y_is and

$$\frac{vS^2}{\sigma^2} \sim \chi^2(v).$$

Tukey's method depends on knowing the distribution of the *Studentized range* when $\mu_1 = \mu_2 = \cdots = \mu_t$, i.e., we need to know that

$$Q \equiv \frac{\max_i y_i - \min_i y_i}{S} \sim Q(t,v)$$

and we need to be able to find percentage points of the $Q(t,v)$ distribution. These are tabled in many books on statistical methods, e.g., Christensen (1996a), Snedecor and Cochran (1980), and Kutner, Nachtsheim, Neter, and Li (2005).

If the observed value of Q is too large, the null hypothesis $H_0 : \mu_1 = \cdots = \mu_t$ should be rejected. That is because any differences in the μ_is will tend to make the range large relative to the distribution of the range when all the μ_is are equal. Since the hypothesis $H_0 : \mu_1 = \cdots = \mu_t$ is equivalent to the hypothesis $H_0 : \mu_i = \mu_j$ for all i and j, we can use the Studentized range test to test all pairs of means. Reject the hypothesis that $H_0 : \mu_i = \mu_j$ if

$$\frac{|y_i - y_j|}{S} > Q(1-\alpha,t,v),$$

where $Q(1-\alpha,t,v)$ is the $(1-\alpha)100$ percentage point of the $Q(t,v)$ distribution. If $H_0 : \mu_i = \mu_j$ for all i and j is true, then at least one of these tests will reject H_0 if and only if

$$\frac{\max_i y_i - \min_i y_i}{S} > Q(1-\alpha,t,v),$$

which happens with probability α. Thus the experimentwise error rate is exactly α.

EXAMPLE 5.4.1. *Two-Way ANOVA.*
Consider the model

$$y_{ijk} = \mu + \alpha_i + \beta_j + e_{ijk}, \quad e_{ijk}\text{s i.i.d. } N(0,\sigma^2),$$

$i = 1,\ldots,a$, $j = 1,\ldots,b$, $k = 1,\ldots,N$. Suppose we want to test the hypotheses $H_0 : \beta_j = \beta_{j'}$ for all $j \neq j'$. Consider the $\bar{y}_{\cdot j \cdot}$ values. Here

$$\bar{y}_{\cdot j \cdot} \sim N\big(\mu + \bar{\alpha}_\cdot + \beta_j, \sigma^2/aN\big)$$

and the $\bar{y}_{\cdot j \cdot}$s are independent because $\bar{y}_{\cdot j \cdot}$ depends only on $\bar{e}_{\cdot j \cdot}$ for its randomness; and since $\bar{e}_{\cdot j \cdot}$ and $\bar{e}_{\cdot j' \cdot}$ are based on disjoint sets of the e_{ijk}s, they must be independent. We will see in Section 7.1 that the $\bar{y}_{\cdot j \cdot}$s are least squares estimates of the $\mu + \bar{\alpha}_\cdot + \beta_j$s, so the $\bar{y}_{\cdot j \cdot}$s must be independent of the MSE. It follows quickly that if $H_0 : \beta_1 = \cdots = \beta_b$ is true, then

$$\frac{\max_j \bar{y}_{\cdot j \cdot} - \min_j \bar{y}_{\cdot j \cdot}}{\sqrt{MSE/aN}} \sim Q(b, dfE);$$

and we reject $H_0 : \beta_j = \beta_{j'}$ if

$$|\bar{y}_{\cdot j \cdot} - \bar{y}_{\cdot j' \cdot}| > Q(1-\alpha,b,dfE)\sqrt{MSE/aN}.$$

Note that Tukey's method provides a competitor to the usual analysis of variance F test for the hypothesis $H_0 : \beta_1 = \cdots = \beta_b$. Also, Tukey's method is only applicable when all the means being used are based on the same number of observations.

5.5 Multiple Range Tests: Newman–Keuls and Duncan

The Newman–Keuls multiple range method is a competitor to the Tukey method. It looks at all pairs of means. In fact, it amounts to a sequential version of Tukey's method. Using the notation of the previous section, order the y_is from smallest to largest, say

$$y_{(1)} \leq y_{(2)} \leq \cdots \leq y_{(t)},$$

and define $\mu_{(i)} = \mu_j$ when $y_{(i)} = y_j$. Note that the $\mu_{(i)}$s need not be ordered in any particular way. However, the Newman–Keuls method acts as if the $\mu_{(i)}$s are also ordered. With this notation, we can write the Studentized range as

$$Q = \frac{y_{(t)} - y_{(1)}}{S}.$$

The Newman–Keuls method rejects $H_0 : \mu_{(t)} = \mu_{(1)}$ if $y_{(t)} - y_{(1)} > SQ(1-\alpha, t, \nu)$. If this hypothesis is not rejected, stop. All means are considered equal. If this hypothesis is rejected, we continue.

The next step tests two hypotheses. $H_0 : \mu_{(t-1)} = \mu_{(1)}$ is rejected if $y_{(t-1)} - y_{(1)} > SQ(1-\alpha, t-1, \nu)$. $H_0 : \mu_{(t)} = \mu_{(2)}$ is rejected if $y_{(t)} - y_{(2)} > SQ(1-\alpha, t-1, \nu)$. If $\mu_{(t-1)} = \mu_{(1)}$ is not rejected, then $\mu_{(1)}, \mu_{(2)}, \ldots, \mu_{(t-1)}$ are assumed to be equal, and no more tests concerning only those means are performed. Similar conclusions hold if $\mu_{(t)} = \mu_{(2)}$ is not rejected. If either hypothesis is rejected, the next round of hypotheses is considered.

The next round of hypotheses includes three hypotheses: $H_0 : \mu_{(t)} = \mu_{(3)}$, $H_0 : \mu_{(t-1)} = \mu_{(2)}$, and $H_0 : \mu_{(t-2)} = \mu_{(1)}$. The hypothesis $H_0 : \mu_{(t-3+i)} = \mu_{(i)}$ is rejected if $y_{(t-3+i)} - y_{(i)} > SQ(1-\alpha, t-2, \nu)$ for $i = 1, 2, 3$.

The procedure continues until, at the last round, the hypotheses $H_0 : \mu_{(i)} = \mu_{(i-1)}$, $i = 2, \ldots, t$, are considered. An hypothesis is rejected if $y_{(i)} = y_{(i-1)} > SQ(1-\alpha, 2, \nu)$.

Remember that if, say, $H_0 : \mu_{(t-1)} = \mu_{(1)}$ is not rejected, we will never test $H_0 : \mu_{(t-1)} = \mu_{(2)}$ or $H_0 : \mu_{(t-2)} = \mu_{(1)}$ in the next round or any other hypothesis in any other round that involves only $\mu_{(1)}, \mu_{(2)}, \ldots, \mu_{(t-1)}$.

The experimentwise error rate is exactly α because if $H_0 : \mu_1 = \cdots = \mu_t$ is true, the Newman–Keuls procedure will conclude that there is a difference in the means if and only if the Tukey method concludes that there is a difference. Because $Q(1-\alpha, 2, \nu) < Q(1-\alpha, 3, \nu) < \cdots < Q(1-\alpha, t-1, \nu) < Q(1-\alpha, t, \nu)$, the Newman–Keuls method will reject the hypothesis that a pair of means is equal more often than Tukey's method. The Newman–Keuls method is thus more powerful. On the

other hand, for pairs of μs that are equal, the Newman–Keuls method will make more mistakes than the Tukey method.

EXAMPLE 5.5.1. Let $\alpha = .01$, $v = 10$, $S = 1$, $t = 5$, and $y_1 = 6.5$, $y_2 = 1.2$, $y_3 = 6.9$, $y_4 = 9.8$, $y_5 = 3.4$. We need the numbers $Q(0.99,5,10) = 6.14$, $Q(0.99,4,10) = 5.77$, $Q(0.99,3,10) = 5.27$, $Q(0.99,2,10) = 4.48$. Ordering the y_is gives

i	2	5	1	3	4
y_i	1.2	3.4	6.5	6.9	9.8

To test $H_0 : \mu_4 = \mu_2$, consider $9.8 - 1.2 = 8.6$, which is larger than $SQ(0.99,5,10) = 6.14$. There is a difference. Next test $H_0 : \mu_2 = \mu_3$. Since $6.9 - 1.2 = 5.7$ is less than 5.77, we conclude that $\mu_2 = \mu_5 = \mu_1 = \mu_3$. We do no more tests concerning only those means. Now test $H_0 : \mu_5 = \mu_4$. Since $9.8 - 3.4 = 6.4 > 5.77$, we reject the hypothesis.

We have concluded that $\mu_2 = \mu_5 = \mu_1 = \mu_3$, so the next allowable test is $H_0 : \mu_1 = \mu_4$. Since $9.8 - 6.5 = 3.4 < 5.27$, we conclude that $\mu_1 = \mu_3 = \mu_4$.

Drawing lines under the values that give no evidence of a difference in means, we can summarize our results as follows:

i	2	5	1	3	4
y_i	1.2	3.4	6.5	6.9	9.8

Note that if we had concluded that $\mu_2 \neq \mu_3$, we could test $H_0 : \mu_2 = \mu_1$. The test would be $6.5 - 1.2 = 5.3 > 5.27$, so we would have rejected the hypothesis. However, since we concluded that $\mu_2 = \mu_5 = \mu_1 = \mu_3$, we never get to do the test of $\mu_2 = \mu_1$.

Duncan has a multiple range test that is similar to Newman–Keuls but where the α levels for the various rounds of tests keep decreasing. In fact, Duncan's method is exactly the same as Newman–Keuls except that the α levels used when taking values from the table of the Studentized range are different. Duncan suggests using a $1 - (1 - \alpha)^{p-1}$ level test when comparing a set of p means. If there is a total of t means to be compared, Duncan's method only controls the experimentwise error rate at $1 - (1 - \alpha)^{t-1}$. For $\alpha = 0.05$ and $t = 6$, Duncan's method can only be said to have an experimentwise error rate of 0.23. As Duncan suggests, his method should only be performed when a corresponding omnibus F test has been found significant. This two stage procedure may be a reasonable compromise between the powers of the LSD and Newman–Keuls methods.

5.6 Summary

The emphasis in this chapter has been on controlling the experimentwise error rate. We have made some mention of power and the fact that increased power can be

a mixed blessing. The really difficult problem for multiple comparison procedures is not in controlling the experimentwise error rate, but in carefully addressing the issues of power and the sizes of individual tests.

The discussion of Duncan's multiple range test highlights an interesting fact about multiple comparison methods. Any method of rejecting hypotheses, if preceded by an appropriate omnibus F test, is a valid multiple comparison procedure, valid in the sense that the experimentwise error rate is controlled. For example, if you do an F test first and stop if the F test does not reject, you can then 1) reject all individual hypotheses if the analysis is being performed on your mother's birth date, 2) reject no individual hypotheses on other dates. As stupid as this is, the experimentwise error rate is controlled. Intelligent choice of a multiple comparison method also involves consideration of the error rates (probabilities of type I errors) for the individual hypotheses. The main question is: If not all of the hypotheses are true, how many of the various kinds of mistakes do the different methods make?

A reasonable goal might be to have the experimentwise error rate and the error rates for the individual hypotheses all no greater than α. The Scheffé, LSD, Bonferroni, Tukey, and Newman–Keuls methods all seem to aim at this goal. The Duncan method does not seem to accept this goal.

Suppose we want α level tests of the hypotheses

$$H_0 : \lambda_k'\beta = 0, \qquad k \in \Omega.$$

A reasonable procedure is to reject an hypothesis if

$$SS(\lambda_k'\beta)/MSE > C$$

for some value C. For example, the LSD method takes $C = F(1-\alpha, 1, dfE)$. If Ω consists of all the hypotheses in a t-dimensional space, Scheffé's method takes $C = tF(1-\alpha, t, dfE)$. If Ω is a finite set, say $\Omega = \{1, \ldots, r\}$, then the Bonferroni method takes $C = F(1-\alpha/r, 1, dfE)$.

To control the level of the individual test $H_0 : \lambda_k'\beta = 0$, one needs to pick C as the appropriate percentile of the distribution of $SS(\lambda_k'\beta)/MSE$. At one extreme, if one ignores everything else that is going on and if $\lambda_k'\beta$ was chosen without reference to the data, the appropriate distribution for $SS(\lambda_k'\beta)/MSE$ is $F(1, dfE)$. At the other extreme, if one picks $\lambda_k'\beta$ so that $SS(\lambda_k'\beta)$ is maximized in a t-dimensional space, then the appropriate distribution for $SS(\lambda_k'\beta)/MSE$ is t times an $F(t, dfE)$; it is clear that the probability of rejecting any hypothesis other than that associated with maximizing $SS(\lambda_k'\beta)$ must be less than α. Thus, in the extremes, we are led to the LSD and Scheffé methods. What one really needs is the distribution of $SS(\lambda_k'\beta)/MSE$ given $\lambda_k'\beta = 0$, and all the information contained in knowing $\lambda_j'\beta$ for $j \in \Omega - \{k\}$ and that $SS(\lambda_j'\beta)$ for $j \in \Omega - \{k\}$ will also be observed. Since the desired distribution will depend on the $\lambda_j'\beta$s, and they will never be known, there is no hope of achieving this goal.

The quality of the LSD method depends on how many hypotheses are to be tested. If only one hypothesis is to be tested, LSD is the method of choice. If all of the hypotheses in a subspace are to be tested, LSD is clearly a bad choice for testing

the hypothesis that maximizes $SS(\lambda_k'\beta)$ and also a bad choice for testing other hypotheses that look likely to be significant. For testing a reasonably small number of hypotheses that were chosen without looking at the data, the LSD method seems to keep the levels of the individual tests near the nominal level of α. (The fact that the individual hypotheses are tested only if the omnibus F test is rejected helps keep the error rates near their nominal levels.) However, as the number of hypotheses to be tested increases, the error rate of the individual tests can increase greatly. The LSD method is not very responsive to the problem of controlling the error level of each individual test, but it is very powerful in detecting hypotheses that are not zero.

Scheffé's method puts an upper bound of α on the probability of type I error for each test, but for an individual hypothesis chosen without examining the data, the true probability of type I error is typically far below α. Scheffé's method controls the type I error but at a great cost in the power of each test.

The Bonferroni method uses the same distributions for $SS(\lambda_k'\beta)/MSE$, $k \in \Omega$, as the LSD method uses. The difference is in the different ways of controlling the experimentwise error rate. Bonferroni reduces the size of each individual test, while LSD uses an overall F test. The Bonferroni method, since it reduces the size of each test, does a better job of controlling the error rate for each individual hypothesis than does the LSD method. This is done at the cost of reducing the power relative to LSD. For a reasonable number of hypotheses, the Bonferroni method tends to be more powerful than Scheffé's method and tends to have error levels nearer the nominal than Scheffé's method.

A similar evaluation can be made of the methods for distinguishing between pairs of means. The methods that are most powerful have the highest error rates for individual hypotheses. From greatest to least power, the methods seem to rank as LSD, Duncan, Newman–Keuls, Tukey. Scheffé's method should rank after Tukey's. The relative position of Bonferroni's method is unclear.

When deciding on a multiple comparison method, one needs to decide on the importance of correctly identifying nonzero hypotheses (high power) relative to the importance of incorrectly identifying zero hypotheses as being nonzero (controlling the type I error). With high power, one will misidentify some zero hypotheses as being nonzero. When controlling the type I error, one is more likely not to identify some nonzero hypotheses as being nonzero.

Table 5.1 contains a summary of the methods considered in this chapter. It lists the hypotheses for which each method is appropriate, the method by which the experimentwise error rate is controlled, and comments on the relative powers and probabilities of type I error (error rates) for testing individual hypotheses.

Exercise 5.3 Show that for testing all hypotheses in a six-dimensional space with 30 degrees of freedom for error, if the subspace F test is omitted and the nominal LSD level is $\alpha = 0.005$, then the true error rate must be less than 0.25.

Hint: Try to find a Scheffé rejection region that is comparable to the LSD rejection region.

Table 5.1 Summary of Multiple Comparison Methods

Method	Hypotheses	Control	Comments
Scheffé	Any and all hypotheses constraining a particular subspace	F test for subspace	Lowest error rate and power of any method. Good for data snooping. HSD better for pairs of means.
LSD	Any and all hypotheses constraining a particular subspace	F test for subspace	Highest error rate and power of any method. Best suited for a finite number of hypotheses.
Bonferroni	Any finite set of hypotheses	Bonferroni inequality	Most flexible method. Often similar to HSD for pairs of means.
Tukey's HSD	All differences between pairs of means	Studentized range test	Lowest error rate and power for pairs of means.
Newman–Keuls	All differences between pairs of means	Studentized range test	Error rate and power intermediate between HSD and Duncan.
Duncan	All differences between pairs of means	Studentized range test or F test	Error rate and power intermediate between Newman–Keuls and LSD.

5.6.1 Fisher Versus Neyman–Pearson

I have *tried* to maintain a Fisherian view towards statistical inference in this edition of the book. However, I think multiple comparison procedures are fundamentally a tool of Neyman–Pearson testing. Fisherian testing is about measuring the evidence against the null model, while Neyman–Pearson testing is about controlling error rates. Controlling the experimentwise error rate seems anti-Fisherian to me.

Fisher is often credited with (blamed for) the LSD method. However, Fisher (1935, Chapter 24) did not worry about the experimentwise error rate when making multiple comparisons using his least significant difference method in analysis of variance. He did, however, worry about drawing inappropriate conclusions by using an invalid null distribution for tests determined by examining the data. In particular, Fisher proposed a Bonferroni correction when comparing the largest and smallest sample means.

If you are going to look at all pairs of means, then the appropriate distribution for comparing the largest and smallest sample means is the Studentized range. It gives the approriate P value. An appropriate P value for other comparisons is difficult, but P values based on the Studentized range should be conservative (larger than the true P value). Similar arguments can be made for other procedures.

5.7 Additional Exercises

Exercise 5.7.1 Compare all pairs of means for the blue jeans exercise of Chapter 4. Use the following methods:

(a) Scheffé's method, $\alpha = 0.01$,
(b) LSD method, $\alpha = 0.01$,
(c) Bonferroni method, $\alpha = 0.012$,
(d) Tukey's HSD method, $\alpha = 0.01$,
(e) Newman–Keuls method, $\alpha = 0.01$.

Exercise 5.7.2 Test whether the four orthogonal contrasts you chose for the blue jeans exercise of Chapter 4 equal zero. Use all of the appropriate multiple comparison methods discussed in this chapter to control the experimentwise error rate at $\alpha = 0.05$ (or thereabouts).

Exercise 5.7.3 Compare all pairs of means in the coat–shirt exercise of Chapter 4. Use all of the appropriate multiple comparison methods discussed in this chapter to control the experimentwise error rate at $\alpha = 0.05$ (or thereabouts).

Exercise 5.7.4 Suppose that in a balanced one-way ANOVA the treatment means $\bar{y}_{1\cdot}, \ldots, \bar{y}_{t\cdot}$ are not independent but have some nondiagonal covariance matrix V. How can Tukey's HSD method be modified to accommodate this situation?

Exercise 5.7.5 For an unbalanced one-way ANOVA, give the contrast coefficients for the contrast whose sum of squares equals the sum of squares for treatments. Show the equality of the sums of squares.