MULTIPLE COMPARISON PROCEDURES

Sharon Cebula

Department of Mathematical Sciences

Montana State University

September 16, 1997

A writing project submitted in partial fulfillment

of the requirements for the degree

Master of Science in Statistics

APPROVAL

of a writing project submitted by

SHARON CEBULA

This writing project has been read by the writing project director and has been found to be satisfactory regarding content, English usage, format, citations, bibliographic style, and consistency, and is ready for submission to the Statistics Faculty.

Sept 16, 1997

Date

Robert J. Boik

Writing Project Director

1 Introduction

A frequent goal in statistical analysis is to compare three or more means. An omnibus ANOVA F-test can be performed to determine if the means are significantly different from one another. A significant F-test, however, only indicates that at least 2 of the means differ. It does not indicate which means differ. The use of multiple comparison procedures (MCP's) provides an in-depth examination of linear combinations of the treatment means. These linear combinations of the treatment means are called contrasts or comparisons. There are numerous multiple comparison procedures available. Choosing the appropriate procedure depends on the data set, the type of comparisons desired, and also on how Type I error is to be controlled. Computer packages such as SAS perform many different multiple comparison procedures. In using these packages, it is important to understand the properties of the procedures so that the most appropriate one can be chosen. MCP's are often grouped into two categories, simultaneous test procedures (Gabriel, 1969), and multi-stage tests. Three procedures in each of these categories will be examined. Of specific interest is the justification of Ryan's procedure, a multi-stage test.

2 Control of the Type I Error Rate

An important issue in performing multiple comparisons is how to control the Type I error rate. A Type I error occurs when a true null hypothesis is rejected. The following discussion on controlling the error rate is based on Kirk (1995) and Toothaker (1993). More advanced treatments of this topic can be found in Miller (1981) and Hochberg (1987).

For a single comparison, the probability of making a Type 1 error is equal to the significance level, α. When more than one comparison is being considered, the error rate can be controlled in several ways.

1. Per Comparison Error Rate: The probability that any individual comparison will incorrectly be declared significant. The significance level per comparison is set to be

$$\alpha_{pc} = \alpha.$$

2. Familywise Error Rate: The probability of at least one Type I error for the comparisons in the family, where a family is defined as some set of comparisons. The set of comparisons in a family are usually chosen in one of two ways.

A family may consist of an *a priori* set of k specific contrasts. If k *independent* comparisons are of interest to the researcher and each comparison is tested at α_{pc}, then

$$\text{P(one or more Type I errors)} = 1 - (1 - \alpha_{pc})^k.$$

This equation can be motivated as follows. First note that

$$\text{P(not making a Type I error for a single comparison)} = 1 - \alpha_{pc}.$$

The multiplication rule for independent events can be applied to the k independent comparisons so that

$$\text{P(not making any Type I errors)} = (1 - \alpha_{pc})^k.$$

Thus,

$$\begin{aligned} \text{P(making one or more Type I errors)} &= 1 - \text{P(not making any Type 1 errors)} \\ &= 1 - (1 - \alpha_{pc})^k. \end{aligned}$$

In practice, tests are often not independent. This occurs for two reasons. First, the numerators may be correlated. Second, even if the numerators are independent, the denominator, the *MSE*, is usually the same for all test statistics. Thus, the tests are not independent. When k dependent test statistics have one degree of freedom in the numerator, the same estimator of σ^2 in the denominator (*MSE*),

and the same critical value for each t (ie, all $\alpha_{pc,i}$ are equal), Sidak (1967) showed

$$\bigcap_{i=1}^{k} \mathrm{P}(|t_i| < w) \geq \bigcap_{i=1}^{k} \mathrm{P}\left(\frac{|z_i|}{\sqrt{MSE}} < w\right)$$

where w is a constant and $z_i \sim \mathrm{N}(0,\sigma^2)$. Kimball (1951) showed

$$\bigcap_{i=1}^{k} \mathrm{P}\left(\frac{|z_i|}{\sqrt{MSE}} < w\right) \geq \prod_{i=1}^{k} \mathrm{P}\left(\frac{|z_i|}{\sqrt{MSE}} < w\right).$$

Thus,

$$\begin{aligned} 1 - \alpha_{fw} &= \bigcap_{i=1}^{k} \mathrm{P}\left(|t_i| < w\right) \\ &\geq \prod_{i=1}^{k} \mathrm{P}\left(\frac{|z_i|}{MSE} < w\right) \\ &= (1 - \alpha_{pc})^k. \end{aligned}$$

Therefore,

$$\alpha_{fw} \leq 1 - (1 - \alpha_{pc})^k.$$

This inequality is sometimes referred to as the Sidak multiplicative inequality. This upperbound on α_{fw} holds for families of independent comparisons as well as for families of dependent comparisons satisfying the conditions above.

An example of a family constructed in this manner is the set of all pairwise comparisons for a given factor. A one-way classification would consist of one family constructed in this manner. A two-way classification would consist of three families constructed in this manner, one family of comparisons for each factor and one family for the interaction between the two factors.

A family may also consist of all possible contrasts among a set of means. The number of possible

contrasts is infinite. An example of a family constructed in this manner is the set of all possible contrasts among the means for one factor in a two-way classification.

3. Error Rate per Family: The expected number of false rejections made in a finite family of comparisons. Note, this error rate is not a probability, it is an expected value. The error rate per family can be computed by

$$\alpha_{pf} = \sum_{i=1}^{k} \alpha_{pc,i},$$

where k is the number of comparisons and $\alpha_{pc,i}$ is the per comparison error rate for the ith comparison.

4. Experimentwise Error Rate: The probability of at least one Type 1 error among all comparisons in the experiment. The experiment may consist of several families. For a one-way classification, the family and the experiment are often used interchangeably to represent the set of contrasts of interest.

5. Error Rate Per Experiment: The expected number of false rejections made in inferences on all comparisons in the experiment.

Choosing the correct conceptual unit for Type 1 error can be difficult. Kirk (1995) suggested that the comparison be used as the conceptual unit for orthogonal comparisons that are planned in advance (*a priori* comparisons). Orthogonal comparisons are comparisons which are nonredundant. Let Ψ_1 and Ψ_2 denote two comparisons and c_{1j} and c_{2j} their respective coefficients, where $j = 1, ..., J$ denote the treatment levels. The two comparisons are orthogonal if:

$$\sum_{j=1}^{J} \frac{c_{1j}c_{2j}}{n_j} = 0,$$

or for equal sample sizes,

$$\sum_{j=1}^{J} c_{1j}c_{2j} = 0.$$

A larger conceptual unit is suggested for *a priori* nonorthogonal comparisons such as the family of pairwise comparisons because nonorthogonal comparisons contain redundant information. If the family of comparisons contains an infinite number of comparisons, then the error rate per family cannot be controlled and the familywise error rate should be used. A further discussion of how to control the Type 1 error can

be found in Hochberg (1987).

3 Test Statistics

Computations for MCP rely on the use of various statistics. Three of the most widely used statistics will be discussed. Let Ψ denote the population comparison $\Psi = \sum_{j=1}^{J} c_j \mu_j$ and $\hat{\Psi} = \sum_{j=1}^{J} c_j \bar{y}_j$ denote the estimate of the population comparison. A t-statistic can be computed for each comparison using the equation

$$t_{\hat{\Psi}} = \frac{\hat{\Psi} - \Psi_o}{\sqrt{MSE \sum_{j=1}^{J} \frac{c_j^2}{n_j}}}$$

where Ψ_o is value of Ψ under the null hypothesis. The Mean Squared Error is computed by dividing SSE by $dfe = N - J$. The equation for SSE is

$$SSE = \sum_{j=1}^{J} \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_j)^2$$

where $\bar{y}_j$ is the average of the observations for treatment j. For the case of equal sample sizes, n_j can be replaced by n.

In most situations, the hypothesis to be tested is

$$H_o : \Psi = 0,$$

and the corresponding test statistic simplifies to

$$t_{\hat{\Psi}} = \frac{\hat{\Psi}}{\sqrt{MSE \sum_{j=1}^{J} \frac{c_j^2}{n_j}}}.$$

If pairwise comparisons are of interest under the hypothesis $\Psi = 0$, the test statistic simplifies to

$$t_{\hat{\Psi}} = \frac{\bar{y}_j - \bar{y}_{j'}}{\sqrt{MSE\left(\frac{1}{n_j} + \frac{1}{n_{j'}}\right)}} \text{ for } j \neq j'. \tag{1}$$

The Studentized range statistic can be computed for pairwise as well as non-pairwise comparisons using the equation

$$q = \frac{\sum_{j=1}^{J} c_j \bar{y}_j}{\left(\sum_{j=1}^{J} \frac{|c_j|}{2}\right)\sqrt{\frac{MSE}{n}}}$$

where equal sample sizes have been assumed. For pairwise comparisons and equal sample sizes, $t = \frac{q}{\sqrt{2}}$. The F statistic is also used by MCP's and is equal to t^2. The statistics used in the MCP's will be presented in terms of the t statistic.

4 Usual t Procedure

The usual t procedure is used to test individual comparisons at α_{pc}. The decision rule is to reject H_o if

$$| t_{\hat{\Psi}} | \geq t_{dfe}^{\alpha_{pc}/2}$$

where $dfe = N - J$ and J = the number of treatment means. An interval estimate of the population comparison is given by

$$\hat{\Psi} \pm t_{dfe}^{\alpha_{pc}/2}\sqrt{MSE\sum_{j=1}^{J}\frac{c_j^2}{n_j}}.$$

For equal sample sizes, substitute n for n_j. This interval also can be used to test the hypothesis $H_o : \Psi = \Psi_o$ by rejecting H_o if the interval contains Ψ_o and failing to reject if it does not. The usual t procedure controls the error rate at $\alpha_{pc} = \alpha$. For any group of $k \geq 2$ comparisons with the same denominator, $\alpha_{fw} \leq 1-(1-\alpha)^k$ which is larger than α.

5 Simultaneous Test Procedures

Three of the MCP's that SAS performs are discussed below. They are special cases of the simultaneous test procedures (STP's) (Gabriel, 1969). In a STP, a single critical value is used for all comparisons. The STP's discussed below can be used to test the two-sided hypothesis $H_o : \Psi = \Psi_o$ by using the decision rule given or by constructing a confidence interval to check for containment of Ψ_o.

5.1 Tukey's Multiple Comparison Procedure

Tukey's procedure (1953) simultaneously controls the probability of one or more Type 1 errors for all comparisons of J means (ie, familywise). This procedure requires equal sample sizes. The decision rule is to reject H_o if

$$| t_{\hat{\Psi}} | \geq \frac{q^{\alpha}_{J,dfe} \sum_{j=1}^{J} |c_j|}{2\sqrt{\sum_{j=1}^{J} c_j^2}},$$

where J = the number of treatment means, $dfe = N - J$, and q is the table value of the Studentized range statistic.

For pairwise comparisons the decision rule reduces to

$$| t_{\hat{\Psi}} | \geq \frac{q^{\alpha}_{J,dfe}}{\sqrt{2}}.$$

If unequal sample sizes are used, the test statistic does not follow the studentized range distribution. In the case of unequal sample sizes, the procedure proposed by Tukey (1953) and Kramer (1956) can be used. This procedure has the same decision rule as above, except the test statistic $t_{\hat{\Psi}}$ proposed for pairwise comparisons and unequal sample sizes (equation 1) is used.

The confidence interval for Ψ associated with Tukey's procedure is

$$\hat{\Psi} \pm \frac{q^{\alpha}_{J,dfe}}{2} \left(\sum_{j=1}^{J} |c_j| \right) \sqrt{\frac{MSE}{n}}.$$

A confidence interval for the difference between two means when the sample sizes are unequal using the Tukey-Kramer modification is

$$\bar{y}_j - \bar{y}_{j'} \pm \frac{q^{\alpha}_{J,dfe}}{\sqrt{2}} \sqrt{MSE\left(\frac{1}{n_j} + \frac{1}{n_{j'}}\right)}.$$

For the case of nonpairwise comparisons, the harmonic mean can be used in place of n (Winer, 1962).

Once again, the confidence interval can be used to test the hypothesis $H_o : \Psi = \Psi_o$ by rejecting H_o if the interval contains Ψ_o and failing to reject if it does not. Tukey's procedure has relatively good power when all pairwise comparisons are being considered. However, it is not as powerful for nonpairwise comparisons as is Scheffé's procedure discussed next.

5.2 Scheffé's Multiple Comparison Procedure

Scheffé's procedure (1953, 1959) controls the error rate familywise for the family of all possible comparisons among a set of means. If comparisons are selected by examining the data ($post - hoc$), an infinite number of comparisons are being examined. Scheffé's procedure is recommended for $post - hoc$ comparisons because it maintains control of the familywise error rate. The decision rule for the Scheffé procedure is to reject H_o if

$$| t_{\hat{\Psi}} | > \sqrt{(J-1)F^{\alpha}_{J-1,dfe}}$$

where J=the number of treatment means, and F is the table value with numerator $df = J - 1$ and denominator $df = dfe$. The confidence interval for Ψ associated with Scheffé's procedure is

$$\hat{\Psi} \pm \sqrt{(J-1)F^{\alpha}_{J-1,dfe}} \sqrt{MSE \sum_{j=1}^{J} \frac{c_j^2}{n_j}}.$$

Scheffé's procedure is relatively powerful when a large number of comparisons are to be made, but is somewhat conservative when a small number of *a priori* comparisons or just pairwise comparisons are of interest.

5.3 Bonferroni/Dunn's Multiple Comparison Procedure

Dunn's procedure (1961) is sometimes referred to as Bonferroni's procedure because it is based on the Bonferroni Inequality. It is appropriate for any set of *a priori* pairwise or nonpairwise comparisons. It controls the Type I error rate per family by dividing the overall significance level into parts. Recall for a family of k comparisons, $\alpha_{pf} = \sum_{i=1}^{k} \alpha_{pc,i}$. If a researcher considers the consequences of making a Type I error to be equally serious for all k contrasts, $\alpha_{pc} = \frac{\alpha_{pf}}{k}$ can be chosen for each comparison. If the consequences of making a Type I error are not equally as serious, α_{pf} can be allocated unequally to represent the concerns of the researcher as described by Kirk (1995). For example, if four comparisons are of interest and committing a Type I error is more serious for the latter two comparisons, the allocation $\alpha_1 = 0.02$, $\alpha_2 = 0.02$, $\alpha_3 = 0.005$, and $\alpha_4 = 0.005$ for the respective comparisons will control the Type I error at $\alpha_{pf} = 0.02 + 0.02 + 0.005 + 0.005 = 0.05$ while taking into account the seriousness of the individual comparisons.

If we consider a family of k independent or dependent *a priori* comparisons, Bonferroni's procedure also controls the Type I error rate familywise as motivated by the following discussion. Let A_i be the event that a Type I error is made on the ith comparison and let A_i^c be the complement of A_i. The probability of not making a Type I error on any of the k independent comparisons can be written as $P\left(\bigcap_{i=1}^{k} A_i^c\right)$. Applying Bonferroni's inequality yields

$$P\left(\bigcap_{i=1}^{k} A_i^c\right) \geq \sum_{i=1}^{k} P(A_i^c) - (k-1).$$

Recall, the probability of not making a Type I error on the ith comparison is equal to $(1 - \alpha_{pc,i})$. Also, the probability of not making a Type I error on any of the k independent comparisons is equal to $1 - \alpha_{fw}$. Therefore,

$$1 - \alpha_{fw} \geq \sum_{i=1}^{k} (1 - \alpha_{pc,i}) - (k-1),$$

which implies,

$$\alpha_{fw} \leq \sum_{i=1}^{k} \alpha_{pc,i},$$

or,

$$\alpha_{fw} \leq k\alpha_{pc}$$

if all $\alpha_{pc,i}$ are equal.

Thus the following relationship exists among the error rates for k test statistics:

$$\alpha_{pc,i} \leq \alpha_{fw} \leq [\alpha_{pf} = \sum_{i=1}^{k} \alpha_{pc,i}].$$

For k dependent test statistics having the same denominator and all $\alpha_{pc,i}$ equal, the Bonferroni additive inequality and Sidak multiplicative inequality yield

$$\alpha_{fw} \leq 1 - (1 - \alpha_{pc})^k \leq [\alpha_{pf} = k\alpha_{pc}].$$

For small levels of α and a reasonable number of comparisons, the familywise error rate and the per family error rate are almost identical.

The decision rule for Bonferroni's procedure is to reject H_o if

$$| t_{\hat{\Psi}} | \geq t_{dfe}^{\alpha_{pc}/2k},$$

where equal $\alpha_{pc,i}$ have been assumed. The confidence interval equation corresponding to Bonferroni's procedure is

$$\hat{\Psi} \pm t_{dfe}^{\alpha_{pc}/2k} \sqrt{MSE \sum_{j=1}^{J} \frac{c_j^2}{n_j}}.$$

Because Bonferroni's procedure relies on the number of comparisons being computed, it is most powerful for a small number of comparisons.

6 Multi-Stage Procedures

Multi-stage procedures use adjusted significance levels in which means of the same stretch size are tested at the same significance level. A stretch size, symbolized by p, is the number of means in the range of the pair of ordered means being tested. For example, the stretch size is 3 when comparing the largest and smallest of 3 means. Although simultaneous test procedures are easier to understand than multi-stage test procedures, the latter are generally more powerful. Multi-stage procedures were developed to test hypotheses rather than to construct confidence intervals. Much of the literature suggests that multi-stage procedures generally cannot be used to compute confidence intervals (Kirk, 1982; Einot and Gabriel, 1975; Toothaker, 1993). Hayter and Hsu (1994) proposed modifications of step-wise procedures which allow for the construction of simultaneous confidence intervals. A number of multi-stage procedures can be performed using SAS. Three of these procedures will be discussed.

6.1 Newman-Keuls Multiple Comparison Procedure

The procedure due to Newman (1939) and Keuls (1952) is performed using the following step-down logic. The J means are arranged in order from smallest to largest. For J treatment means, the stretch sizes vary from $p = J$, the set of all means, to $p = 2$, the subset of two adjacent means. First conduct a comparison test of the largest and smallest mean. This is the comparison test for stretch size $p = J$. If the comparison test for stretch size $p = J$ is not significant, retain H_o and all hypotheses for stretch sizes $p \leq J$. If the comparison test for stretch size $p = J$ is significant, reject H_o and proceed testing the two comparisons of stretch size $p = J - 1$. In other words, test the smallest and second largest mean and test the largest and second smallest mean at α_{J-1}. Repeat this process until conclusions for comparison tests for $p = 2$ have been made. The Newman-Keuls procedure controls the Type I error at $\alpha_p = \alpha$ for stretch size p. Under the complete null hypothesis, equality of all J means, Newman-Keuls procedure controls the familywise Type I error rate at α. Under a partial null hypothesis, the Newman-Keuls procedure may not control the familywise Type I error rate at α. If the number of treatments is less than or equal to three, the Type I error is controlled familywise. If $J \geq 3$, α_{fw} may exceed α. Consider the case $J = 4$ with two sets of equal

means such that $\mu_1 = \mu_2 < \mu_3 = \mu_4$. A test of $\mu_1 = \mu_2$ will be performed at α, and a test of $\mu_3 = \mu_4$ will be tested at α. Thus, α_{fw} is bounded by $1-(1-\alpha)^2 = \alpha(2-\alpha)$ which is greater than α.

The decision rule is to reject H_o if

(i) the means in the comparison are not contained in the stretch of a previously retained hypothesis, and

(ii) $|\, t_{\hat{\Psi}} \,| \geq \frac{q^{\alpha}_{p,df_e}}{\sqrt{2}}$.

Einot and Gabriel (1975) argued that Newman-Keuls procedure is unsatisfactory because the Type I error rate may not be controlled familywise. A modification of the procedure was proposed that would control the error rate experimentwise at α. However, this modification involves complicated computations with only a slight increase in power and is not recommended (Einot and Gabriel, 1975).

6.2 Duncan's Multiple Comparison Procedure

Duncan's MCP (1955) follows the same step-down logic of Newman-Keuls procedure. Duncan's procedure, however, controls the error rate at $\alpha_p = 1-(1-\alpha)^{p-1}$ for stretch size p. As p increases, the test is designed to achieve greater power. This increase in power is due to an inflated familywise error rate. For $p \geq 3$, the Type I error rate is not controlled familywise. For this reason Duncan's procedure is not recommended unless one is willing to compromise a relaxed Type I error rate for an increase in power. Ryan's procedure, which will be discussed next, is usually preferred to both Newman-Keuls and Duncan's procedure.

6.3 Ryan's Procedure (REGWQ)

Ryan (1960) proposed a multi-stage procedure which utilizes the same step-down logic as Newman-Keuls and Duncan's procedures, but does control the familywise Type I error rate at α. Ryan suggested conducting a comparison test for stretch size $p = J$, a test of the largest and smallest mean, by using a critical value from the t-distribution with $\alpha_J = \frac{2\alpha}{J(J-1)}$. If the test is not significant, retain H_o and all hypotheses for stretch sizes $p \leq J$. If the test is significant, conduct the two comparison tests for stretch size $p' = J-1$ using the t critical value with significance level $\frac{2\alpha}{J(J-2)}$. Proceed as discussed in the step-down logic of the Neuman-Keuls procedure using two-sample t-tests with significance level $\frac{2\alpha}{J(p-1)}$ for stretch size p. Because

there are $\frac{p(p-1)}{2}$ pairwise comparisons for stretch size p, the Type I error rate for stretch size p does not exceed

$$\begin{aligned}\alpha_p &= \frac{2\alpha}{J(p-1)}\frac{p(p-1)}{2} \\ &= \frac{p\alpha}{J}.\end{aligned}$$

The motivation behind Ryan's procedure follows. When the complete null hypothesis is true (ie, when all of the J treatment means are equal) and all $\frac{J(J-1)}{2}$ pairwise comparison tests are conducted at $\frac{2\alpha}{J(J-1)}$,

$$\alpha_{pf} = \frac{J(J-1)}{2}\frac{2\alpha}{J(J-1)} = \alpha.$$

For the family of all comparisons, $\alpha_{fw} \leq \alpha_{pf}$. Therefore, by testing the extreme pair at $\frac{2\alpha}{J(J-1)}$, the Type I error rate will be controlled familywise at α when all treatment means are equal. If the complete hypothesis is false, then at least 2 treatment means differ. Partition the means into G sets of homogeneous means. A Type I error can be made only within a homogeneous set of means. Let n_g denote the number of means in set g where $\sum_{g=1}^{G} n_g = J$. If the $\frac{n_g(n_g-1)}{2}$ pairwise comparisons in group g are each tested at $\frac{2\alpha}{J(n_g-1)}$, an upperbound on the probability of one or more Type I errors is given by the expected number of Type I errors; namely

$$\frac{n_g(n_g-1)}{2}\frac{2\alpha}{J(n_g-1)} = \frac{n_g}{J}\alpha.$$

Therefore, the probability of not making a Type I error for group g is no smaller than $(1-\frac{n_g}{J}\alpha)$. Ignoring the common denominator (acting as though the sets of means are independent), the familywise Type I error rate for the G sets of contrasts is

$$\alpha_{fw} \approx 1 - \prod_{g=1}^{G}(1-\frac{n_g}{J}\alpha)$$

which Ryan (1960) showed is less than or equal to

$$\frac{\sum_{g=1}^{G} n_g}{J}\alpha = \frac{J\alpha}{J} = \alpha.$$

Thus, Ryan's procedure controls the familywise Type I error rate at α regardless of how the means are partitioned into homogeneous groups. If the g^{th} set of homogeneous means is tested using an F test with $n_g - 1$ numerator degrees of freedom, then Kimball's (1951) result shows that

$$\alpha_{fw} \leq 1 - \prod_{g=1}^{G}(1 - \frac{n_g}{J}\alpha).$$

Ryan's procedure was modified by Einot and Gabriel (1975) who proposed using $1 - (1-\alpha)^{\frac{n_g}{J}}$ as an upperbound on the probability of one or more Type I errors for group g rather than $\frac{n_g}{J}\alpha$ as in Ryan's procedure. Following the argument above,

$$\begin{aligned}
\alpha_{fw} &\approx 1 - \prod_{g=1}^{G}(1 - [1 - (1-\alpha)^{\frac{n_g}{J}}]) \\
&= 1 - (1-\alpha)^{\sum_{g=1}^{G}\frac{n_g}{J}} \\
&= 1 - (1-\alpha)^{\frac{J}{J}} \\
&= \alpha.
\end{aligned}$$

Thus, Einot and Gabriel's modification controls the error rate familywise at α by using $\alpha_p = 1 - (1-\alpha)^{\frac{p}{J}}$ as an upperbound on the Type I error rate for the comparisons of stretch size p. This modification of Ryan's procedure improves the test's power.

Ryan's initial proposal was based on the two-sample t-test. Welsch (1977) proposed utilizing the more powerful Studentized range distribution with $\alpha_p = 1 - (1-\alpha)^{\frac{p}{J}}$ for stretch sizes $2 \leq p \leq J-2$ and $\alpha_p = \alpha$ for stretch sizes $p = J-1$ and $p = J$. When testing the means within stretch size $p = J-1$, a Type I error can only be made within the $J-1$ means. If $\alpha_{J-1} = 1 - (1-\alpha)^{\frac{J-1}{J}}$ is used for stretch size $J-1$, a portion of the familywise error rate, $1 - (1-\alpha)^{\frac{1}{J}}$, is being allotted to the single mean from a given population. However, it is impossible to make a Type I error when there is only one mean. Thus using $\alpha_{J-1} = \alpha$ will increase the power for this stretch size while controlling the familywise error rate at α.

The resulting procedure is often referred to as Ryan's procedure in the literature, but SAS has given credit to all three authors by calling it REGWF or REGWQ, depending on whether the F-statistic or the Range

statistic is used, where the letters REGW refer to Ryan, Einot, Gabriel, and Welsch. The REGWF procedure has the advantage of being compatible with the overall F-test, but is more computationally demanding. The decision rule for the REGWQ procedure is to reject H_o if

(i) the means in the comparison are not contained in the stretch of a previously retained hypothesis, and

(ii) $|t_{\bar{\psi}}| \geq \frac{q_{p,df_e}^{\alpha_p}}{\sqrt{2}}$.

where

$$\alpha_p = \begin{cases} \alpha & \text{for } p = J \text{ or } p = J-1; \text{ and} \\ 1-(1-\alpha)^{\frac{p}{J}} & \text{for } 2 \leq p \leq J-2. \end{cases}$$

7 Numerical Examples

The following data on location strategies for gypsy moth scent-lure traps is taken from a text by Lapin (1990).

Trap Location Strategy				
(1)	(2)	(3)	(4)	(5)
Scattered	Concentrated	Host Plant	Aerial	Ground
90	99	95	98	87
92	97	96	98	93
94	98	97	99	90
93	98	97	99	91
-	99	96	-	89

The response variable is the estimated percentage of the native male population trapped.

SAS was used to perform all pairwise comparisons at $\alpha = 0.05$. The SAS program is listed in the appendix.

7.1 Simultaneous Test Procedures

For the STP's discussed, it is important to use the CLDIFF option in SAS to display the results. If this option is not used, SAS uses the harmonic mean for n in the case of unequal sample sizes. The tests performed in this manner will not be exact. Using the CLDIFF option will display the exact results based on unequal sample sizes for the Scheffé and Dunn procedures, and will use the Tukey-Kramer procedure in place of Tukey's procedure.

7.1.1 Tukey's Procedure

The following output was produced by SAS.

General Linear Models Procedure

Tukey's Studentized Range (HSD) Test for variable: PERCT

NOTE: This test controls the type I experimentwise error rate.

Alpha= 0.05 Confidence= 0.95 df= 18 MSE= 1.963889

Critical Value of Studentized Range= 4.276

Comparisons significant at the 0.05 level are indicated by '***'.

LOC Comparison	Simultaneous Lower Confidence Limit	Difference Between Means	Simultaneous Upper Confidence Limit	
4 - 2	-2.5426	0.3000	3.1426	
4 - 3	-0.5426	2.3000	5.1426	
4 - 1	3.2536	6.2500	9.2464	***
4 - 5	5.6574	8.5000	11.3426	***
2 - 4	-3.1426	-0.3000	2.5426	
2 - 3	-0.6800	2.0000	4.6800	

2	- 1	3.1074	5.9500	8.7926	***
2	- 5	5.5200	8.2000	10.8800	***
3	- 4	-5.1426	-2.3000	0.5426	
3	- 2	-4.6800	-2.0000	0.6800	
3	- 1	1.1074	3.9500	6.7926	***
3	- 5	3.5200	6.2000	8.8800	***
1	- 4	-9.2464	-6.2500	-3.2536	***
1	- 2	-8.7926	-5.9500	-3.1074	***
1	- 3	-6.7926	-3.9500	-1.1074	***
1	- 5	-0.5926	2.2500	5.0926	
5	- 4	-11.3426	-8.5000	-5.6574	***
5	- 2	-10.8800	-8.2000	-5.5200	***
5	- 3	-8.8800	-6.2000	-3.5200	***
5	- 1	-5.0926	-2.2500	0.5926	

Recall, the confidence interval formula for $\mu_j - \mu_{j'}$ is

$$\hat{\Psi} \pm \frac{q^{\alpha}_{J,df_E}}{\sqrt{2}} \sqrt{MSE\left(\frac{1}{n_j} + \frac{1}{n_{j'}}\right)}.$$

Using $q^{0.05}_{5,18} = 4.276$, $MSE = 1.963889$, $n_4 = 4$, and $n_2 = 5$, the confidence interval for $\mu_2 - \mu_4$ is 0.3 ± 2.845. This produced the interval $(-2.5426\ ,\ 3.1426)$ above. The other confidence intervals were formed in a similar manner.

The pair of means for an interval which does not contain zero are concluded to be significantly different. From the results above, the means for locations 1 and 5 are found to be significantly different from the means of locations 2, 3, and 4. In addition to being able to test the hypothesis of equal means, the confidence intervals provide an estimate of the true difference between the population means.

Tukey's procedure controls the familywise Type I error rate at $\alpha = 0.05$.

7.1.2 Scheffé's Procedure

The following output was produced using SAS.

Scheffe's test for variable: PERCT

NOTE: This test controls the type I experimentwise error rate but generally has a higher type II error rate than Tukey's for all pairwise comparisons.

Alpha= 0.05 Confidence= 0.95 df= 18 MSE= 1.963889

Critical Value of F= 2.92774

Comparisons significant at the 0.05 level are indicated by '***'.

LOC Comparison	Simultaneous Lower Confidence Limit	Difference Between Means	Simultaneous Upper Confidence Limit	
4 - 2	-2.9171	0.3000	3.5171	
4 - 3	-0.9171	2.3000	5.5171	
4 - 1	2.8589	6.2500	9.6411	***
4 - 5	5.2829	8.5000	11.7171	***
2 - 4	-3.5171	-0.3000	2.9171	
2 - 3	-1.0331	2.0000	5.0331	
2 - 1	2.7329	5.9500	9.1671	***
2 - 5	5.1669	8.2000	11.2331	***
3 - 4	-5.5171	-2.3000	0.9171	
3 - 2	-5.0331	-2.0000	1.0331	

3	- 1	0.7329	3.9500	7.1671	***
3	- 5	3.1669	6.2000	9.2331	***
1	- 4	-9.6411	-6.2500	-2.8589	***
1	- 2	-9.1671	-5.9500	-2.7329	***
1	- 3	-7.1671	-3.9500	-0.7329	***
1	- 5	-0.9671	2.2500	5.4671	
5	- 4	-11.7171	-8.5000	-5.2829	***
5	- 2	-11.2331	-8.2000	-5.1669	***
5	- 3	-9.2331	-6.2000	-3.1669	***
5	- 1	-5.4671	-2.2500	0.9671	

Recall, the confidence interval formula for $\mu_j - \mu_{j'}$ is

$$\hat{\Psi} \pm \sqrt{(J-1)F^{\alpha}_{J-1,dfe}}\sqrt{MSE\left(\frac{1}{n_j}+\frac{1}{n_{j'}}\right)}.$$

Using , $F^{0.05}_{4,18} = 2.92774$, $MSE = 1.963889$, $n_4 = 4$, and $n_2 = 5$, the confidence interval for $\mu_2 - \mu_4$ is 0.3 ± 3.2171. This produced the interval $(-2.9171\ ,\ 3.5171)$ above. The other confidence intervals were formed in a similar manner.

The pair of means for an interval which does not contain zero are concluded to be significantly different. From the results above, the means for locations 1 and 5 are found to be significantly different from the means of locations 2, 3, and 4.

Scheffé's procedure also controls the Type I error familywise at $\alpha = 0.05$.

7.1.3 Bonferroni/ Dunn's Procedure

The following output was produced by SAS. The *a priori* set of interest is all pairwise comparisons.

Bonferroni (Dunn) T tests for variable: PERCT

NOTE: This test controls the type I experimentwise error rate but generally has a higher type II error rate than Tukey's for all pairwise comparisons.

Alpha= 0.05 Confidence= 0.95 df= 18 MSE= 1.963889

Critical Value of T= 3.19657

Comparisons significant at the 0.05 level are indicated by '***'.

LOC Comparison	Simultaneous Lower Confidence Limit	Difference Between Means	Simultaneous Upper Confidence Limit	
4 - 2	-2.7050	0.3000	3.3050	
4 - 3	-0.7050	2.3000	5.3050	
4 - 1	3.0824	6.2500	9.4176	***
4 - 5	5.4950	8.5000	11.5050	***
2 - 4	-3.3050	-0.3000	2.7050	
2 - 3	-0.8332	2.0000	4.8332	
2 - 1	2.9450	5.9500	8.9550	***
2 - 5	5.3668	8.2000	11.0332	***
3 - 4	-5.3050	-2.3000	0.7050	
3 - 2	-4.8332	-2.0000	0.8332	
3 - 1	0.9450	3.9500	6.9550	***
3 - 5	3.3668	6.2000	9.0332	***
1 - 4	-9.4176	-6.2500	-3.0824	***
1 - 2	-8.9550	-5.9500	-2.9450	***
1 - 3	-6.9550	-3.9500	-0.9450	***
1 - 5	-0.7550	2.2500	5.2550	

5	- 4	-11.5050	-8.5000	-5.4950	***
5	- 2	-11.0332	-8.2000	-5.3668	***
5	- 3	-9.0332	-6.2000	-3.3668	***
5	- 1	-5.2550	-2.2500	0.7550	

Recall, the confidence interval formula for $\mu_j - \mu_{j'}$ is

$$\hat{\Psi} \pm t_{dfe}^{\alpha_{FC}/2k} \sqrt{MSE\left(\frac{1}{n_j} + \frac{1}{n_{j'}}\right)}.$$

Using , $t_{18}^{0.05/2(10)} = 3.1966$, $MSE = 1.963889$, $n_4 = 4$, and $n_2 = 5$, the confidence interval for $\mu_2 - \mu_4$ is 0.3 ± 3.0050. This produced the interval $(-2.7050\ ,\ 3.3050)$ above. The other confidence intervals were formed in a similar manner.

The pair of means for an interval which does not contain zero are concluded to be significantly different. From the results above, the means for locations 1 and 5 are found to be significantly different from the means of locations 2, 3, and 4.

Bonferroni's procedure also controls the Type I error familywise at $\alpha = 0.05$.

7.1.4 Comments

Notice that for pairwise comparisons, the Tukey-Kramer procedure produces the shortest confidence intervals with a margin of error for $\mu_4 - \mu_2$ of 2.845. The margin of error for this comparison using Scheffé's and Bonferroni's procedures are 3.2171 and 3.0050 respectively. For all pairwise comparisons, Tukey's procedure is the most powerful of these three tests.

Now suppose the contrast $\frac{\mu_4+\mu_5}{2} - \mu_3$ was selected *a priori* to be examined. The harmonic mean, n_h, will be used in place of n in the construction of the confidence interval based on Tukey's method.

$$\begin{aligned} n_h &= \frac{J}{\sum_{i=1}^{J} \frac{1}{n_j}} \\ &= \frac{5}{\frac{1}{4}+\frac{1}{5}+\frac{1}{5}+\frac{1}{4}+\frac{1}{5}} \\ &= 4.545455. \end{aligned}$$

A 95% confidence interval using Tukey's procedure is

$$\hat{\Psi} \pm \frac{q^{\alpha}_{J,dfe}}{2} \left(\sum_{j=1}^{J} |c_j| \right) \sqrt{\frac{MSE}{n_h}},$$

which simplifies to

$$\begin{aligned} \frac{98.5+90}{2} - 96.2 &\pm (\frac{4.276}{2})(\frac{1}{2}+\frac{1}{2}+1)\sqrt{\frac{1.963889}{4.545455}} \\ -1.95 &\pm 2.8107 \end{aligned}$$

$$(-4.7607\,,\, 0.86065).$$

A 95% confidence interval using Scheffé's procedure is

$$\hat{\Psi} \pm \sqrt{(J-1)F^{\alpha}_{J-1,dfe}} \sqrt{MSE \sum_{j=1}^{J} \frac{c_j^2}{n_j}},$$

which simplifies to

$$\begin{aligned} -1.95 &\pm \sqrt{4(2.92774)}\sqrt{1.963889\left(\frac{1}{16}+\frac{1}{20}+\frac{1}{5}\right)} \\ -1.95 &\pm 2.6809 \end{aligned}$$

$$(-4.6309\,,\,0.73089).$$

A 95% confidence interval using Bonferroni's procedure is

$$\hat{\Psi} \pm t_{df_e}^{\alpha_{pc}/2k} \sqrt{MSE \sum_{j=1}^{J} \frac{c_j^2}{n_j}}.$$

If $k = 1$ and $C = (0\ 0\ -1\ \frac{1}{2}\ \frac{1}{2})'$ is the vector of c_j's, then the confidence interval is as follows:

$$\begin{aligned} -1.95 &\pm 2.101\sqrt{1.963889\left(\frac{1}{16}+\frac{1}{20}+\frac{1}{5}\right)} \\ -1.95 &\pm 1.646 \end{aligned}$$

$$(-3.5959\,,\,-0.304).$$

Thus for a *single* nonpairwise comparison, Bonferroni's procedure produces the smallest interval, followed by Scheffé and Tukey respectively. In general, Scheffé's procedure is more powerful than Tukey's when nonpairwise comparisons are of interest. For a small number of comparisons Bonferroni's procedure tends to be more powerful than both.

7.2 Multi-stage Tests

The multi-stage procedures are designed for only pairwise comparisons and assume equal sample sizes. For the case of unequal sample sizes, SAS substitutes the harmonic mean n_h for n. Montgomery (1991) also suggested this substitution in his presentation of Duncan's procedure. The harmonic mean of the sample sizes is $n_h = 4.545455$.

7.2.1 Newman-Keuls Procedure

The following output was obtained from SAS.

Student-Newman-Keuls test for variable: PERCT

NOTE: This test controls the type I experimentwise error rate under the complete null hypothesis but not under partial null hypotheses.

Alpha= 0.05 df= 18 MSE= 1.963889

WARNING: Cell sizes are not equal.

Harmonic Mean of cell sizes= 4.545455

Number of Means	2	3	4	5
Critical Range	1.9529654	2.3724293	2.6272522	2.8108484

Means with the same letter are not significantly different.

SNK Grouping	Mean	N	LOC
A	98.5000	4	4
A	98.2000	5	2
A	96.2000	5	3
B	92.2500	4	1
C	90.0000	5	5

The procedure discussed would reject $H_o : \mu_j - \mu_{j'} = 0$ if

$$| t_{\bar{y}_j - \bar{y}_{j'}} | \geq \frac{q^{\alpha}_{p,N-J}}{\sqrt{2}}$$

for stretch size p. SAS chooses to use an equivalent criteria and rejects H_o if

$$| \bar{y}_j - \bar{y}_{j'} | \geq \frac{q^{\alpha}_{p,N-J}}{\sqrt{2}} \sqrt{MSE\left(\frac{1}{n_h} + \frac{1}{n_h}\right)}.$$

The results given by SAS were produced in the following manner. The sample means for the locations were first ordered.

Location	5	1	3	2	4
$\bar{y}$	90.00	92.25	96.20	98.20	98.50

For stretch size $p = 5$, the hypothesis $H_o : \mu_4 - \mu_5 = 0$ is tested at $\alpha = 0.05$. The critical value for $p = 5$ is

$$\begin{aligned} \frac{q^{0.05}_{5,18}}{\sqrt{2}} &= \frac{4.28}{\sqrt{2}} \\ &= 3.0264. \end{aligned}$$

The test statistic is

$$| t_{\bar{y}_4 - \bar{y}_5} | = 9.1440.$$

Because $9.1440 > 3.0264$, the null hypothesis is rejected. This is equivalent to comparing $| \bar{y}_4 - \bar{y}_5 | = 8.5$ to

$$\begin{aligned} 3.90264\left(\sqrt{1.963889\left(\frac{2}{4.545455}\right)}\right) &= 3.90264(0.92958) \\ &= 2.8108, \end{aligned}$$

the critical value reported by SAS.

Next, the two comparisons for stretch size $p = 4$ are tested.

The critical value for $p = 4$ used by SAS is

$$\frac{q^{0.05}_{4,18}}{\sqrt{2}}\sqrt{1.963889\left(\frac{2}{4.545455}\right)} = \frac{4.00}{\sqrt{2}}(0.92958)$$

$$= \quad 2.6273.$$

The test statistic for the hypothesis $H_o : \mu_5 - \mu_2 = 0$ used by SAS is

$$| \bar{y}_5 - \bar{y}_2 | = 8.20.$$

Because $8.2 > 2.6273$ the null hypothesis is rejected.

The test statistic for the hypothesis $H_o : \mu_1 - \mu_4 = 0$ is

$$| \bar{y}_1 - \bar{y}_4 | = 6.25.$$

Because $6.25 > 2.6273$ the null hypothesis is rejected.

The critical value used for the comparisons of stretch size $p = 3$ is

$$\frac{q^{0.05}_{3,18}}{\sqrt{2}}\sqrt{1.963889\left(\frac{2}{4.545455}\right)} = \frac{3.61}{\sqrt{2}}(0.92958)$$
$$= \quad 2.3274.$$

The test statistic for the hypothesis $H_o : \mu_5 - \mu_3 = 0$ is

$$| \bar{y}_5 - \bar{y}_3 | = 6.20.$$

Because $6.20 > 2.3274$ the null hypothesis is rejected.

The test statistic for the hypothesis $H_o : \mu_1 - \mu_2 = 0$ is

$$| \bar{y}_1 - \bar{y}_2 | = 5.95.$$

Because $5.95 > 2.3274$, the null hypothesis is rejected.

The test statistic for the hypothesis $H_o : \mu_3 - \mu_4 = 0$ is

$$| \bar{y}_3 - \bar{y}_4 | = 2.30.$$

Because $2.30 \not> 2.3274$, the null hypothesis is retained. In addition, the hypotheses $H_o : \mu_3 - \mu_2 = 0$ and $H_o : \mu_2 - \mu_4 = 0$ are also retained.

Lastly, test the hypothesis $H_o : \mu_5 - \mu_1 = 0$ for stretch size $p = 2$.

The critical value used for $p = 2$ is

$$\begin{aligned} \frac{q^{0.05}_{2,18}}{\sqrt{2}}\sqrt{1.963889\left(\frac{2}{4.545455}\right)} &= \frac{2.97}{\sqrt{2}}(0.92958) \\ &= 1.9530. \end{aligned}$$

The test statistic is 2.25. Since $2.25 > 1.9530$, the null hypotheis is rejected.

Newman-Keuls procedure found locations 1 and 5 to be significantly different from the other 3 locations and from each other.

7.2.2 Duncan's Procedure

The following output was produced by SAS.

```
        Duncan's Multiple Range Test for variable: PERCT
NOTE: This test controls the type I comparisonwise error rate,
      not the experimentwise error rate
          Alpha= 0.05  df= 18  MSE= 1.963889
          WARNING: Cell sizes are not equal.
       Harmonic Mean of cell sizes= 4.545455
     Number of Means      2      3      4      5
     Critical Range   1.953  2.049  2.110  2.152
```

Means with the same letter are not significantly different.

Duncan Grouping	Mean	N	LOC
A	98.5000	4	4
A	98.2000	5	2
B	96.2000	5	3
C	92.2500	4	1
D	90.0000	5	5

The results were found using the same process as Newman-Keuls, but $q^{\alpha}_{p,N-J}$ is replaced with $q^{1-(1-\alpha)^{p-1}}_{p,N-J}$. The results are displayed in the table below. The Studentized range values $q^{0.1855}_{5,18} = 3.274$, $q^{0.1426}_{4,18} = 3.210$, $q^{0.0975}_{3,18} = 3.118$, and $q^{0.05}_{2,18} = 2.971$ were used in the computations of the critical values.

H_o	p	Test Statistic (SAS)	Critical Value	Decision
$\mu_4 - \mu_5 = 0$	5	8.5	$\frac{3.274}{\sqrt{2}}(0.92958) = 2.152$	Reject H_o
$\mu_5 - \mu_2 = 0$	4	8.20	$\frac{3.210}{\sqrt{2}}(0.92958) = 2.110$	Reject H_o
$\mu_1 - \mu_4 = 0$	4	6.25	2.110	Reject H_o
$\mu_5 - \mu_3 = 0$	3	6.20	$\frac{3.118}{\sqrt{2}}(0.92958) = 2.049$	Reject H_o
$\mu_1 - \mu_2 = 0$	3	5.95	2.049	Reject H_o
$\mu_3 - \mu_4 = 0$	3	2.30	2.049	Reject H_o
$\mu_5 - \mu_1 = 0$	2	2.25	$\frac{2.971}{\sqrt{2}}(0.92958) = 1.953$	Reject H_o
$\mu_1 - \mu_3 = 0$	2	3.95	1.953	Reject H_o
$\mu_3 - \mu_2 = 0$	2	2.00	1.953	Reject H_o
$\mu_2 - \mu_4 = 0$	2	0.30	1.953	Fail to Reject H_o

Duncan's procedure found all pairs of locations means to be significantly different from one another except for locations 2 and 4.

Duncan's procedure does not control the familywise Type I error rate at $\alpha = 0.05$.

7.2.3 REGWQ

The following output was obtained from SAS

Ryan-Einot-Gabriel-Welsch Multiple Range Test for variable: PERCT

NOTE: This test controls the type I experimentwise error rate.

Alpha= 0.05 df= 18 MSE= 1.963889

WARNING: Cell sizes are not equal.

Harmonic Mean of cell sizes= 4.545455

Number of Means	2	3	4	5
Critical Range	2.3658807	2.6027605	2.6272522	2.8108484

Means with the same letter are not significantly different.

REGWQ Grouping	Mean	N	LOC
A	98.5000	4	4
A	98.2000	5	2
A	96.2000	5	3
B	92.2500	4	1
B	90.0000	5	5

The results were found using the same process as Newman-Keuls for $p = 4$ and $p = 5$. For $p \leq 3$, $q^{\alpha}_{p,N-J}$ is replaced with $q^{1-(1-\alpha)^{\frac{p}{J}}}_{p,N-J}$. The results are displayed in the table below. The Studentized Range values $q^{0.05}_{5,18} = 4.28$, $q^{0.05}_{4,18} = 4.00$, $q^{0.0303}_{3,18} = 3.960$, and $q^{0.0203}_{2,18} = 3.599$ were used in the computations of the critical values.

H_o	p	Test Statistic (SAS)	Critical Value	Decision
$\mu_4 - \mu_5 = 0$	5	8.5	$\frac{4.28}{\sqrt{2}}(0.92958) = 2.1808$	Reject H_o
$\mu_5 - \mu_2 = 0$	4	8.20	$\frac{4.00}{\sqrt{2}}(0.92958) = 2.6273$	Reject H_o
$\mu_1 - \mu_4 = 0$	4	6.25	2.6273	Reject H_o
$\mu_5 - \mu_3 = 0$	3	6.20	$\frac{3.960}{\sqrt{2}}(0.92958) = 2.6028$	Reject H_o
$\mu_1 - \mu_2 = 0$	3	5.95	2.6028	Reject H_o
$\mu_3 - \mu_4 = 0$	3	2.30	2.6028	Fail to Reject H_o
$\mu_5 - \mu_1 = 0$	2	2.25	$\frac{3.599}{\sqrt{2}}(0.92958) = 2.3659$	Fail to Reject H_o

7.2.4 Comments

The following table compares the results using Newman-Keuls, Duncan's, and REGWQ. Means with the same letter are not significantly different.

SNK	Duncan	REGWQ	Mean	N	Loc
A	A	A	98.50	4	4
A	A	A	98.20	5	2
A	B	A	96.20	5	3
B	C	B	92.25	4	1
C	D	B	90.00	5	5

Newman-Keuls and Duncan's procedures detected a larger number of significantly different means than Ryan's procedure did. However, neither Newman-Keuls procedure nor Duncan's procedure controls the familywise Type I error rate at $\alpha = 0.05$. In fact, Duncan's procedure is even more liberal with α_{fw} than Newman-Keuls procedure. Thus the detections of $\mu_1 \neq \mu_5$ by Newman-Keuls and $\mu_1 \neq \mu_5$, $\mu_3 \neq \mu_1$ by Duncan's are potential Type I errors.

8 Conclusion

Finding an appropriate MCP can be difficult. An important consideration in choosing a MCP is how the Type I error rate is controlled. It is usually desirable to control the familywise Type I error rate at a specified level α. For more than two comparisons, the usual t procedure does not maintain familywise control of the error rate. The simultaneous test procedures discussed, Tukey, Scheffé, and Bonferroni, do maintain familywise control of the Type I error. In addition, these procedures allow for the use of simultaneous confidence statements about the means. Confidence intervals can not only be used to make decisions regarding hypothesis tests of the means, but they also provide an estimate of the true differences between population means. There has been a general movement towards relying less on significance tests in reporting results, and instead utilizing more confidence intervals. Schmidt (1996) commented, "Use of point estimates of effect size and confidence intervals in interpreting data in individual studies would have made our research literatures far less confusing, far less apparently contradictory, and far more informative than those that have been produced by the dominant practice of reliance on significance tests". Schmidt (1996) also refers to other supporters of this reasoning.

Ryan's procedure maintains familywise control of the Type I error rate and is said to be even more powerful than Tukey's procedure for pairwise comparisons. However, Ryan's procedure is designed for only pairwise comparisons and has not been developed for the use of confidence intervals. If other contrasts are of interest, or simultaneous confidence bounds are desired, an appropriate simultaneous test procedure should be selected.

9 References

1. Duncan, D. B. (1955), "Multiple Range and Multiple F Tests," *Biometrics*, 11, 1-42.

2. Dunn, O. J. (1961), "Multiple Comparisons Among Means," *Journal of the American Statistical Association*, 56: 52-64.

3. Einot, I. and Gabriel, K. R. (1975), "A Study of the Powers of Several Methods of Multiple Compar-

isons," *Journal of the American Statistial Association*, 70, 351.

4. Freund, R. J., Littell, R. C., and Spector, P. C. (1986), *SAS System for Linear Models, 1986 Edition*, Cary, NC: SAS Institute Inc.

5. Hayter, A. J., Hsu, J. C. (1994), "On the Relationship Between Stepwise Decision Procedures and Confidence Sets," *Journal of the American Statistical Association*, 89, 128-136.

6. Kirk, R. E. (1982), *Experimental Design: Procedures for the Behavioral Sciences*, Belmont, CA: Brookes/Cole.

7. Keuls, M. (1952), "The Use of the 'Studentized Range' in Connection with an Analysis of Variance," *Euphytica*, 1: 112-122.

8. Kimball, A.W., "On Dependent Tests of Significance in the Analysis of Variance," *Annals of Mathematical Statistics*, 22: 600-602.

9. Kramer, C. Y. (1956), "Extension of Multiple Range Tests to Group Means with Unequal Numbers of Replications," *Biometrics*, 12: 307-310.

10. Lapin, Lawrence L. (1990), *Probability and Statistics for Modern Engineering*, Boston: PWS-Kent.

11. Miller, R. G., Jr. (1981), *Simultaneous Statistical Inference*, New York: Springer-Verlag.

12. Montgomery, Douglas C., (1991), *Design and Analysis of Experiments*, New York: John Wiley & Sons, Inc.

13. Newman, D. (1939), "The Distribution of the Range in Samples from a Normal Population, Expressed in Terms of an Independent Estimate of Standard Deviation," *Biometrika*, 31: 20-30.

14. Ryan, T. A. (1959), "Multiple Comparisons in Psychological Research," *Psychological Bulletin*, 56: 26-47.

15. Ryan, T. A. (1960), "Significance Tests for Multiple Comparison of Proportions, Variance, and Other Statistics," *Psychological Bulletin*, 57: 318-328.

16. Scheffe, H. (1953), "A Method for Judging All Contrasts in the Analysis of Variance," *Biometrika*, 40, 87-104.

17. Scheffe, H. (1959), *The Analysis of Variance*, New York: John Wiley & Sons, Inc.

18. Schmidt, Frank L. (1996), "Statistical Significance Testing and Cumulative Knowledge in Psychology: Implications for Training of Researchers," *Psychological Bulletin*, 1: 115-129.

19. Sidak, Z. (1967), "Rectangular Confidence Regions for the Means of Multivariate Normal Distributions," *Journal of the American Statistical Association*, 62, 626-633.

20. Toothaker, L. E. (1993) *Multiple Comparison Procedures* (Sage University Paper series on Quantitative Applications in the Social Sciences, series no 07-089). Newbury Park, CA: Sage.

21. Tukey, J. W. (1953), "The Problem of Multiple Comparisons," Unpublished manuscript.

22. Welsch, R. E. (1977), "Stepwise Multiple Comparison Procedures", *Journal of the American Statistical Association*, 72: 566-575.

23. Winer, B. J. (1962), *Statistical Principles in Experimental Design*, New York: McGraw-Hill.

10 Appendix

```
data in;
input loc perct@@;
cards;
 1 90  1 92  1 94  1 93
 2 99  2 97  2 98  2 98  2 99
 3 95  3 96  3 97  3 97  3 96
 4 98  4 98  4 99  4 99
 5 87  5 93  5 90  5 91  5 89
;
```

```
proc glm data=in;

class loc;

model perct=loc /ss3;

means loc / cldiff bon tukey scheffe;

means loc /  duncan snk regwq;

run;
```