SIAM REVIEW
Vol. 44, No. 2, pp. 259–278

Multiple Comparison Methods for Means*

John A. Rafter†
Martha L. Abell†
James P. Braselton†

Abstract. Multiple comparison methods (MCMs) are used to investigate differences between pairs of population means or, more generally, between subsets of population means using sample data. Although several such methods are commonly available in statistical software packages, users may be poorly informed about the appropriate method(s) to use and/or the correct way to interpret the results. This paper classifies the MCMs and presents the important methods for each class. Both simulated and real data are used to compare methods, and emphasis is placed on correct application and interpretation. We include suggestions for choosing the best method.

Mathematica programs developed by the authors are used to compare MCMs. By taking advantage of *Mathematica*'s notebook structure, an interested student can use these programs to explore the subject more deeply. The programs and examples used in the article are available at http://www.cs.gasou.edu/faculty/rafter/MCMM/.

Key words. multiple comparison procedures, familywise error rate, single-step procedures, step-down procedures

AMS subject classification. 62

PII. S0036144501357233

1. Introduction.

A common problem in the sciences and industry is to compare several treatments to determine which, if any, produce a superior outcome. For example, suppose a manufacturer wants to examine the effect on sales due to package design. A reasonable way to proceed is to select a group of stores with comparable sales volumes and randomly and independently assign each store to carry one of the package designs to be tested. Assume several stores carry each package design and conditions that could affect sales, such as price, shelf space, and promotional efforts, are the same for all stores [18].

When the data gathering is finished, it may turn out that one package design is clearly superior to the others. In this case there is no need for statistical analysis of the data. On the other hand, the average or mean sales for each design may be close enough that it is not easy to decide whether their differences are real or are due to the inherent variation in sales among the stores. A common method for investigating such differences is called *analysis of variance*, often abbreviated to ANOVA.

*Received by the editors June 16, 1999; accepted for publication (in revised form) June 25, 2001; published electronically May 1, 2002.
http://www.siam.org/journals/sirev/44-2/35723.html

†Department of Mathematics and Computer Science, Georgia Southern University, Statesboro, GA 30460-8093 (jarafter@gsaix2.cc.gasou.edu, somatla@gsvms2.cc.gasou.edu, jimbras@gsvms2.cc.gasou.edu).

To make things more precise, assume that there are n stores, there are $k \geq 3$ package designs, and package design i is assigned to n_i stores, where $\sum_{i=1}^{k} n_i = n$. Further, let μ_i be the unknown mean sales that would result if package design i is chosen. Then we can formulate a statistical hypothesis test to look for differences among these (population) means using the sample data from the n stores. The *null hypothesis* is

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k, \tag{1.1}$$

which represents the assertion that all of the means (treatments) are the same. In our example, H_0 states that the package designs have the same impact on sales. We will call it the *overall null hypothesis*. When the researcher's interest is in whether there are differences among the (population) means, the alternative hypothesis is

$$H_a : \text{not all means are equal.} \tag{1.2}$$

The statistical technique used in this case is known as single-factor ANOVA. It is also called an F-test, because the calculation results in a number (called, in general, a *test statistic*) denoted by F. Based on the value of F, the decision is made to either reject or not reject the overall null hypothesis. When the decision is to reject the null hypothesis, the inference made is that there is some difference among the population means. Multiple comparison methods (MCMs) are designed to investigate differences between specific pairs of means or linear combinations of means. This provides the information that is of most use to the researcher.

1.1. Per-Comparison Approach. One possible approach to the multiple comparison problem is to make each comparison independently using a suitable statistical procedure. For example, a statistical hypothesis test could be used to compare each pair of means, μ_i and μ_j, where the null and alternative hypotheses are of the form

$$\begin{aligned} H_0 &: \mu_i = \mu_j, \\ H_a &: \mu_i \neq \mu_j. \end{aligned} \tag{1.3}$$

The usual statistical technique in this case is known as the t-test, so named because the test statistic (and its probability distribution) is denoted by t. With this test, as with any hypothesis test, there is a chance of making errors. One possible error is to reject the null hypothesis, concluding that μ_i and μ_j are different, when in fact they are equal and the null hypothesis is true. The other possible error is to accept the null hypothesis, concluding that μ_i and μ_j are equal, when in fact they are different. These errors are called *Type I error* and *Type II error*, respectively. Any rule for deciding between H_0 and H_a is assessed in terms of probabilities of the two types of errors. If, for example, the decision is to reject H_0, then the probability of a Type I error should be very small. We denote this probability by the Greek letter α and define it using the probability operator P, as

$$\alpha = P\left(\text{reject } H_0 \text{ given } H_0 \text{ is true}\right). \tag{1.4}$$

The quantity α is also called the *level of significance*. By specifying a level of significance (a commonly used value is $\alpha = 0.05$) for the t-test, the researcher controls the probability of finding an erroneous difference. When each of several hypothesis tests is done at the same level of significance α, then α is called the *per-comparison level of significance*.

An alternative way to test for a difference between μ_i and μ_j is to calculate a confidence interval for $\mu_i - \mu_j$. A confidence interval is formed using a point estimate, a margin of error, and the formula

$$\text{(point estimate)} \pm \text{(margin of error)}. \tag{1.5}$$

The point estimate is the best guess for the value of $\mu_i - \mu_j$ based on the sample data. In our example, it would be the difference in mean sales for package designs i and j. The margin of error reflects the accuracy of the guess based on variability in the data. It also depends on a confidence coefficient, which is often denoted by $1-\alpha$. The interval is calculated by subtracting the margin of error from the point estimate to get the lower limit and adding the margin of error to the point estimate to get the upper limit. A commonly used value for the confidence coefficient is $1 - \alpha = 0.95$, and the interval is called a 95% confidence interval. The confidence coefficient is an expression of how certain we are that our experimental procedure will result in an interval that contains $\mu_i - \mu_j$.

The alternative to the t-test for a difference between μ_i and μ_j at level of significance α is to calculate the confidence interval for $\mu_i - \mu_j$ with confidence coefficient $1-\alpha$. If the interval does not contain zero (thereby ruling out that $\mu_i = \mu_j$), then the null hypothesis is rejected and μ_i and μ_j are declared different at level of significance α. For several confidence intervals, each with confidence coefficient $1 - \alpha$, the $1 - \alpha$ is called the *per-comparison confidence coefficient*.

The difficulty with the per-comparison approach to multiple comparisons is the possible inflation of the overall probability of Type I error or (equivalently) the possible deflation of the overall confidence level. For example, consider two independent hypothesis tests, each done at level of significance α. Then the probability that neither has a Type I error is $(1-\alpha)^2$. In other words, the probability of at least one Type I error is $1-(1-\alpha)^2$. In general, to test for differences between every pair of k means would require a total of $c = \frac{1}{2}k(k-1)$ t-tests at per-comparison level of significance α. So the chance of finding at least one erroneous difference is $\alpha_c = 1-(1-\alpha)^c$. For $k \geq 3$, not only is α_c larger than α, but it approaches 1 as k increases. Practically speaking, if you insist on performing many pairwise comparisons at per-comparison level of significance α, you are almost certain to conclude that two of the treatments are different even though they are not.

1.2. Family. A family is a collection of inferences for which it is meaningful to take into account some overall measure of errors. For example, the collection of all pairwise comparisons just discussed is a family, where the overall measure of errors is the probability of encountering at least one Type I error. This family is an example of a finite family (containing c elements). A family may also have an infinite number of members. For example, inferences that incorporate every contrast in the set of all contrasts of k means would form an infinite family. A contrast is a linear combination of two or more means where the coefficients sum to zero. They are discussed in more detail in section 2.3.

In the context of multiple comparisons, the choice of family members depends to a great extent on the type of research being conducted: confirmatory or exploratory. In confirmatory research, it is typical to specify a finite number of inferences prior to the study. These inferences may be unrelated and could be treated separately. For example, ANOVA can be used to separately test for selected differences. But, if there is to be some joint measure of the probability of making erroneous decisions, then these inferences should be considered jointly as a family. In exploratory research,

the comparisons are selected after looking at the data. In order to account for this selection, the family must include both the inferences that are actually made and those that potentially could have been made if suggested by the data. Thus the family may be infinite. Most research is made up of aspects of both approaches, and the corresponding families include the comparisons that were prespecified as well as the comparisons that were suggested by the data.

In the package design example, suppose that designs 1 and 2 are in three colors and designs 3 and 4 are in five colors. Then a finite family could have been prespecified to include, for example, all pairwise comparisons and a comparison of three colors versus five colors. The latter comparison would incorporate a contrast to compare a combination of μ_1 and μ_2 versus a combination of μ_3 and μ_4. On the other hand, if it was decided to compare three colors versus five colors after beginning the analysis, the family must be the infinite family that contains not only all pairwise comparisons but all other possible contrasts involving the means.

1.3. Error Rates. The concept of error rate is most appropriately applied to hypothesis testing, where it represents the probability of Type I error or equivalently the level of significance. As discussed earlier, when a family consists of several hypothesis tests and each hypothesis test is done at the same level of significance α, then α is the *per-comparison error rate* (PCER), the probability of incorrectly rejecting *each* of the null hypotheses that make up the family. A per-comparison confidence level is similarly defined. A more pertinent error rate is called the *familywise error rate* (FWER) or *familywise level of significance*. It is the probability of incorrectly rejecting *at least one* of the null hypotheses that make up the family. The *familywise confidence coefficient* can be similarly defined. It is the chance that the sampling will result in confidence intervals that simultaneously contain the specified combinations of the means in the family.

Having specified an FWER, the researcher must be careful to perform a multiple comparison analysis that guarantees the error rate will be valid under all possible configurations of the population means. Such an analysis will be said to protect the FWER. Consider the family of all pairwise comparisons of a set of k means. A multiple comparison analysis originally proposed by Fisher [11, 16] consists of first doing an F-test. If the overall null hypothesis (1.1) is rejected, at level of significance α, then hypothesis tests for differences between pairs of population means are done using separate t-tests at PCER α. If the overall null hypothesis is not rejected, then the analysis terminates. This procedure is known as Fisher's least significant difference (LSD) test. The familywise level of significance is α only when $\mu_1 = \mu_2 = \cdots = \mu_k$. For any other configuration of the population means (in which they are somehow different), the familywise level of significance is greater than α. Thus, the LSD test does not protect the FWER.

A second multiple comparison analysis, also proposed by Fisher [11, 16] and commonly called the Bonferroni test, consists of performing a t-test for each pair of means at PCER $\alpha/\left(\frac{1}{2}k(k-1)\right)$. Using this test, the familywise level of significance is at most α for any configuration of the population means. Thus, the Bonferroni test protects the FWER. This approach to multiple comparisons using the Bonferroni test illustrates a third type of error rate known as the *per-family error rate* (PFER). It is not a probability, as the other error rates are, but represents the expected number of errors in the family. For example, assuming the overall null hypothesis to be true, if each of c tests is performed with probability of Type I error α/c, the expected number of Type I errors is $c \times (\alpha/c) = \alpha$. Thus, when $\mu_1 = \mu_2 = \cdots = \mu_k$, the PFER is α.

For any other configuration of the means, the PFER would be less than α. It is a straightforward and worthwhile exercise to verify (see [16, p. 8], [20, p. 12]) that

$$\text{PCER} \leq \text{FWER} \leq \text{PFER}. \tag{1.6}$$

The focus of this article is on the familiar and not so familiar MCMs that protect the FWER. They are certainly appropriate when a conclusion is reached that requires the simultaneous correctness of several inferences. For an indication of when strategies other than protecting the FWER might be appropriate, see Dunnett and Tamhane [9] or Hochberg and Tamhane [16].

1.4. MCMs. Statistical procedures that are designed to take into account and control the inflation of the overall probability of Type I error or the deflation of the overall confidence coefficient are called MCMs. These MCMs can be categorized as either single-step or stepwise. For single-step methods, each of the inferences is carried out in a single step without reference to the other inferences in the family. Single-step MCMs that protect the FWER include the Tukey procedure for equal sample sizes (i.e., balanced designs) and the Tukey–Kramer procedure for unequal sample sizes (i.e., unbalanced designs); the Dunnett procedure; the Scheffe procedure; procedures based on approximations like those of Bonferroni and Šidák (pronounced "shedoc"); and procedures appropriate when the population variances are not all equal, such as the C procedure (for Cochran) and the T3 procedure (for Tamhane).

Stepwise procedures make comparisons in a series of steps, where the results on the current step influence which, if any, comparisons are made on the next step. They can be divided into two types: step-down and step-up.

A step-down procedure may begin, for example, by testing the overall null hypothesis (1.1) and, if it is rejected, move to the next step. On succeeding steps a null hypothesis is tested for a subset of means only if they were part of a larger set of means for which a null hypothesis was rejected during an earlier step. Fisher's LSD test is an example of a two-step, step-down procedure. The widely used procedures known as the Newman–Keuls and the Duncan multiple range tests are examples of step-down procedures. These three procedures are *not* recommended, because they do not protect the FWER. A procedure due to Ryan [19] that was modified first by Einot and Gabriel [10] and then by Welsch [21] and a procedure due to Peritz are step-down procedures that do protect the FWER.

A step-up procedure may begin, for example, by testing a pairwise hypothesis (1.3) and, depending on the results, step up to a hypothesis involving a larger number of means. At each succeeding step a decision is made to involve more means or to stop. There are relatively few step-up procedures in the literature, beginning with one described in 1977 by Welsch [21]. Two more recently proposed procedures are in papers by Hochberg [15] and Dunnett and Tamhane [9].

1.5. Assumptions. All inferential procedures have associated assumptions to be satisfied by the data in order to obtain valid results. The MCMs for population means, as well as the F-test, have the same assumptions.

1. The samples are randomly and independently selected.
2. The populations are normally distributed.
3. The populations all have the same variance.

Because one or more of these assumptions may be violated for a given set of data, it is important to be aware of how this would impact an inferential procedure. The insensitivity of a procedure to one or more violations of its underlying assumptions is

called its *robustness*. The first assumption is the least likely to be violated, because it is under the control of the researcher. If violated, neither the MCMs nor the F-test are robust. Most MCMs seem to be robust under moderate departures from normality of the underlying populations in that the actual FWER will only be slightly higher than specified. Some MCMs have been specifically developed to be used when the population variances are not all equal.

1.6. Author-Written Procedures. The programs used to demonstrate the MCMs were written by the authors in *Mathematica*. They are contained in *Mathematica* notebooks, which are available on the web at

`http://www.cs.gasou.edu/faculty/rafter/MCMM/`.

The calculations for the examples in this paper are also contained in *Mathematica* notebooks at the web site. For someone who has access to *Mathematica* but lacks familiarity, the web site contains a series of eight integrated laboratory exercises that introduce *Mathematica* notebooks and demonstrate the use of *Mathematica* for data simulation and for statistical data analysis. All of the notebooks were developed using *Mathematica* on a Power Macintosh but can be implemented in *Mathematica* on any platform.

Mathematica should not be viewed as generally competing with special-purpose statistical packages, although it can be easily used for a variety of statistical analyses in the absence of such a package [1]. For the investigation of MCMs, its strength lies in the notebook structure. For example, the results of several competing MCMs are easily displayed together; changes are easily made and procedures quickly reexecuted; new output is easily displayed along with earlier output, or it can replace earlier output; and simulations are particularly easy to set up.

1.7. Notation. The following notation is used consistently throughout the rest of the article. The sample means for k groups are denoted by $\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_k$, and the respective sample sizes by $n_1, n_2, \ldots, n_k$. The estimate of the common population variance (see assumption (3)) is denoted by s^2. The symbol ν, which is often called degrees of freedom, represents the value $N - k$ (i.e., $\nu = N - k$), where N is the total number of data values for the k groups. The familywise level of significance specified by the researcher is denoted by α, and the corresponding familywise confidence coefficient is denoted by $1 - \alpha$.

Quantiles are important in the calculations for multiple comparisons, so they are introduced here. Consider first a specific probability distribution, which describes the variable D. Then, for any α, where $0 < \alpha < 1$, the quantile corresponding to α is the number d_α that satisfies

$$P(D \leq d_\alpha) = 1 - \alpha.$$

For example, $d_{0.05}$ represents the 95th quantile. When calculating a confidence interval, the quantile d_α corresponding to the confidence coefficient $1 - \alpha$ is used to calculate the margin of error. In the formulas that we present in this paper, we use the convention of including the parameters of the probability distribution as subscripts along with α in the notation for the quantile.

2. Single-Step Procedures.

2.1. Introduction. The common practice of employing an MCM only after performing an F-test on the overall null hypothesis (1.1) is theoretically unnecessary. It continues presumably because the ANOVA calculations provide numerical results

used in the multiple comparison calculations. An experimenter who must reject the overall null hypothesis before trying one or more of the single-step MCMs is using a two-step procedure and may miss an important result. A brief discussion of the negative implications of using such two-step procedures is given in Hochburg and Tamhane [16, pp. 108–109]. The MCMs discussed in this article are appropriately used without reference to the F-test.

Using Confidence Intervals. Single-step MCMs are valid to use both for hypothesis testing and to calculate confidence intervals. Consider a set of k means and the family of all pairwise comparisons containing $c = \frac{1}{2}k(k-1)$ hypothesis tests of the form (1.3). Assuming the familywise level of significance is α, this family is equivalent to a family of c confidence intervals of the form

$$L_{ij} \le \mu_i - \mu_j \le U_{ij}, \tag{2.1}$$

where the familywise confidence coefficient is $1-\alpha$. The equivalence lies in the fact that when the confidence interval contains zero (i.e., 0 lies between L_{ij} and U_{ij}), the null hypothesis, $H_0 : \mu_i = \mu_j$, would not be rejected, but when the interval does not contain zero, the null hypothesis would be rejected and μ_i and μ_j declared to be different at familywise level of significance α. In the latter case, there is an advantage to using a confidence interval, because it provides information about the size and the direction of the difference. With familywise confidence coefficient $1-\alpha$, the difference $\mu_i - \mu_j$ being between L_{ij} and U_{ij} means that, based on the data, the difference may be as small as L_{ij} or as large as U_{ij}. Moreover, if L_{ij} is positive, then $\mu_i > \mu_j$, or if U_{ij} is negative, then $\mu_i < \mu_j$.

The hypothesis test of the form (1.3) is known as a two-sided test, because the form of the alternative hypothesis, H_a, is $\mu_i \neq \mu_j$. When the form of the alternative hypothesis is either $\mu_i < \mu_j$ or $\mu_i > \mu_j$, the test is said to be one-sided. The equivalence of hypothesis testing with confidence intervals does not hold as just described for one-sided hypothesis tests unless special one-sided confidence intervals are calculated. These are discussed in section 2.4.

Comparing MCMs. The *power* of a hypothesis test is a measure of its ability to identify differences. Because identifying differences is usually the reason for the analysis, an appropriate hypothesis test with the highest power is preferred. Thus, to compare among appropriate MCMs, one should look at their ability to identify differences. When the analysis uses confidence intervals, the MCM having the *shortest intervals* is the most powerful.

2.2. Pairwise Comparisons. In this subsection we discuss single-step MCMs for families of pairwise comparisons. For the MCMs that assume equal population variances, the choice of the most powerful depends on the number of comparisons in the family, and there are really only two choices. For the family of all pairwise comparisons use the Tukey (or Tukey–Kramer) test. For a family containing some, but not all, of the pairwise comparisons, use one of, in order of preference, the GT2 procedure, the Dunn–Šidák test, or the Bonferroni test when it gives shorter confidence intervals than the Tukey (or Tukey–Kramer) test. Among the MCMs appropriate when not all population variances are equal, the choice depends on whether the user wants to be completely certain to protect the FWER and sample size. All of the MCMs discussed are used for two-sided hypothesis tests.

Tukey Test (Pairs). The Tukey test is also known as both Tukey's honestly significant difference (HSD) test and Tukey's wholly significant difference (WSD) test. It is an exact test in that, for the family of all $c = \frac{1}{2}k(k-1)$ pairwise comparisons,

the FWER is exactly α (and the familywise confidence coefficient is exactly $1-\alpha$). Exact MCMs are rare. The Tukey test has been shown analytically to be optimal in the sense that, among all procedures that give equal-width confidence intervals for all pairwise differences with familywise confidence coefficient at least $1-\alpha$, it gives the shortest intervals [12]. In effect, this says that, if the family consists of all pairwise comparisons and the Tukey test can be used, it will have shorter confidence intervals than any of the other single-step MCMs.

Given an FWER of α, the Tukey confidence interval for $\mu_i - \mu_j$ is calculated using the formula

$$\bar{y}_i - \bar{y}_j \pm q_{\alpha,\nu,k}\sqrt{\frac{s^2}{n}}, \tag{2.2}$$

where $q_{\alpha,\nu,k}$ represents the quantile for the Studentized range probability distribution with parameters ν and k. The margin of error in (2.2) (the quantity to the right of the $\pm$ sign) is the same for each pairwise comparison, so the confidence intervals will all have the same width. It depends on the total number of means k, so it is not affected by the number of comparisons in the family. The margin of error also depends on the common sample size, n. The one problem with the Tukey test is that it requires a balanced design (i.e., $n_1 = n_2 = \cdots = n_k = n$).

Tukey–Kramer Test (Pairs). The Tukey–Kramer test is a modification of the Tukey test for unbalanced designs. It is not exact but is minimally conservative in that the actual FWER is often only slightly less than α [7, 14]. The confidence interval for $\mu_i - \mu_j$ is calculated using the formula

$$\bar{y}_i - \bar{y}_j \pm q_{\alpha,\nu,k}\sqrt{\frac{1}{2}s^2\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}. \tag{2.3}$$

It has been confirmed analytically that, for unbalanced designs, this formula provides uniformly shorter intervals than any of the other common single-step MCMs [16, p. 106] for the family of all pairwise comparisons.

Bonferroni Test (Pairs). The Bonferroni test can be used for either balanced or unbalanced designs. It is not exact, being based on an approximation known as the first-order Bonferroni inequality. As noted earlier, it is conservative in that the actual FWER is less than α except when the overall null hypothesis (1.1) is true. For the family of all pairwise comparisons, it will produce longer confidence intervals than the Tukey (or Tukey–Kramer) test. Its potential usefulness is in confirmatory research when a family of selected pairwise comparisons is specified prior to data collection. Depending on the size of the family, it may give shorter confidence intervals than the Tukey (or Tukey–Kramer) test. It follows that, if the statistical software for the Bonferroni test does not permit the user to input a specific set of pairwise comparisons, it is of no practical use.

Given an FWER of α, the confidence interval for $\mu_i - \mu_j$ is calculated using the formula

$$\bar{y}_i - \bar{y}_j \pm t_{\alpha',\nu}\sqrt{s^2\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}, \tag{2.4}$$

where $\alpha' = \frac{1}{2}(\alpha/c)$ and c is the number of pairwise comparisons in the family. The quantile $t_{\alpha',\nu}$ is from Student's t probability distribution with parameter ν. Thus the margin of error in (2.4) depends on the number of comparisons.

Dunn–Šidák Test (Pairs). This test was proposed by Dunn [5] and is based on an inequality proved by Šidák. Because the inequality is a slight improvement on the Bonferroni inequality, the confidence intervals will always be a little shorter than those for the Bonferroni test. If available, the Dunn–Šidák test should always be used in place of the Bonferroni test. If a statistical software package offers the GT2 procedure, which was developed by Hochberg based on a further improved inequality also proved by Šidák, it should be used in place of the Dunn–Šidák test (or the Bonferroni test). Like the Bonferroni test, the potential usefulness of the GT2 procedure (the Dunn–Šidák test) is in confirmatory research when a family of selected pairwise comparisons is specified prior to data collection. Depending on the size of the family, it may give shorter confidence intervals than the Tukey (or Tukey–Kramer) test. Given an FWER of α, the Dunn–Šidák confidence interval for $\mu_i - \mu_j$ is calculated using the formula

$$\bar{y}_i - \bar{y}_j \pm t_{\alpha',\nu}\sqrt{s^2\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}, \tag{2.5}$$

where

$$\alpha' = \frac{1}{2}\left(1-(1-\alpha)^{1/c}\right)$$

and c is the number of pairwise comparisons in the family. The quantile $t_{\alpha',\nu}$ is from Student's t probability distribution with parameter ν.

Scheffe Test (Pairs). The Scheffe test is discussed in more detail in section 2.2. It can be used for the family of all pairwise comparisons but will always give longer confidence intervals than the other tests in this subsection. Given an FWER of α, the confidence interval for $\mu_i - \mu_j$ is calculated using the formula

$$\bar{y}_i - \bar{y}_j \pm \sqrt{(k-1)F_{\alpha,k-1,\nu}}\sqrt{s^2\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}, \tag{2.6}$$

where the quantile $F_{\alpha,k-1,\nu}$ is from the F probability distribution with parameters $k-1$ and ν.

Example 1 (pairwise comparisons using single-step MCMs). Four samples of sizes 10, 10, 10, and 20, respectively, were simulated using *Mathematica*. The means of the populations sampled were 5, 6, 6, and 6, respectively, and the variance was 1 for each population. Then each of the Tukey–Kramer, the Bonferroni, the Dunn–Šidák, and the Scheffe single-step MCMs was computed for the family of all pairwise comparisons using $\alpha = 0.05$. The Tukey–Kramer test results are reproduced in Table 1.

Based on the simulation used to generate the data, we know that $\mu_1 < \mu_2 = \mu_3 = \mu_4$. The results show that the Tukey–Kramer test is not able to completely describe this relationship. It does find that $\mu_1 \neq \mu_2$ and $\mu_1 \neq \mu_4$ at the familywise level of significance, $0.05 = 5\%$. The output further shows, with familywise confidence level $0.95 = 95\%$, that $\mu_1 - \mu_2$ is between -2.75 and -0.11 and that $\mu_1 - \mu_4$ is between -2.32 and -0.03, from which we can further conclude that $\mu_1 < \mu_2$ and $\mu_1 < \mu_4$. These are examples of directional decisions. For the single-step MCMs, the chance of making errors in directional decisions is included in the specified FWER along with the chance of making errors in decisions concerning differences. That is, the chance of finding at least one erroneous difference or erroneous direction is α.

The Bonferroni test and the Dunn–Šidák test are only able to show $\mu_1 < \mu_2$, and the Scheffe test finds no statistically significant differences at the 5% familywise level

Table I *Multiple pairwise comparisons using the Tukey–Kramer procedure.*

i	j	$\bar{y}_i - \bar{y}_j$	Interval	Significant difference
1	2	-1.434	$-2.752 \leq \mu_1 - \mu_2 \leq -0.115$	Yes
1	3	-0.720	$-2.039 \leq \mu_1 - \mu_3 \leq 0.598$	No
2	3	0.713	$-0.606 \leq \mu_2 - \mu_3 \leq 2.032$	No
1	4	-1.174	$-2.316 \leq \mu_1 - \mu_4 \leq -0.032$	Yes
2	4	0.259	$-0.883 \leq \mu_2 - \mu_4 \leq 1.402$	No
3	4	-0.454	$-1.596 \leq \mu_3 - \mu_4 \leq 0.689$	No
The familywise confidence level is at least 95%.				
The familywise level of significance is at most 5%.				

of significance. The relative power of the four MCMs is illustrated by the increasing length of the confidence intervals for $\mu_1 - \mu_2$: Tukey–Kramer, 2.64; Dunn–Šidák, 2.72; Bonferroni, 2.73; Scheffe, 2.87. The complete set of commands and output for Example 1 are available at the web site referenced earlier. The commands can be run repeatedly to generate simulated samples from the populations used here and easily modified to generate simulated samples from different populations. Although the pattern of significant differences may change, the lengths of the respective confidence intervals will always have the same ordering. As suggested by the current simulation, differences in interval length between GT2 (not shown), Dunn–Šidák, and Bonferroni MCMs are relatively small.

Procedures for Unequal Variances. There are MCMs that do not require equal population variances and are appropriate for unbalanced designs. The GH procedure due to Games and Howell [13] and the C procedure (for Cochran) due to Dunnett [8] use the Studentized range distribution, where the calculation for the parameter (known as the *degrees of freedom*) includes sample variances. The T3 (for Tamhane) also due to Dunnett [8] uses the Studentized maximum modulus distribution, where the degrees of freedom calculation includes sample variances. The GH procedure is sometimes slightly liberal (the actual FWER is higher than the α specified by the user). In other words, for some data sets, the GH procedure does not protect the FWER. Simulation studies suggest that this may occur, for example, for unequal sample sizes and equal population variances. For one such simulation, the actual FWER was slightly over 0.08 when the user-specified FWER was 0.05. On the other hand, for some data sets, the GH procedure does protect the FWER but is slightly conservative (the actual FWER is lower than the α specified by the user). Simulation studies suggest that this occurs as the population variances diverge. For still other data sets the GH procedure does protect the FWER at the user-specified value of α. At least one source recommends that all sample sizes should be six or more when using the GH procedure.

The two other choices of single-step MCMs that do not require equal population variances, the C and T3 procedures, always protect the FWER. If a researcher is concerned that the GH procedure might be slightly liberal for a given data set, then either the C procedure or the T3 procedure could be used. In this circumstance, the T3 procedure is more powerful than the C procedure for small sample sizes, with the opposite being true for large sample sizes. On the other hand, the researcher may not be concerned about using the GH procedure, either because the population variances are known to be different or because a slight increase in the actual FWER is not a problem. In this circumstance, the GH procedure should be used, because it

is more powerful than both the C and T3 procedures. That is, for any data set, the GH procedure will always have shorter confidence intervals than both the C and T3 procedures.

2.3. Procedures for Contrasts. A *contrast* is a linear combination whose coefficients sum to zero. For example, consider the sample means $\bar{y}_1$, $\bar{y}_2$, $\bar{y}_3$, and $\bar{y}_4$. Because $\bar{y}_2 - \bar{y}_4$ can be written as

$$(0) \cdot \bar{y}_1 + (1) \cdot \bar{y}_2 + (0) \cdot \bar{y}_3 + (-1) \cdot \bar{y}_4$$

and the sum of the coefficients of this linear combination is zero, $\bar{y}_2 - \bar{y}_4$ is a contrast. Similarly, all pairwise differences in means are contrasts. Contrasts are used to make more sophisticated comparisons as well. For example, in the package design example, as suggested earlier, suppose that designs 1 and 2 are in three colors and designs 3 and 4 are in five colors. Then the following null hypothesis may be used to test for the effect of color:

$$H_0 : \frac{1}{2}(\mu_1 + \mu_2) - \frac{1}{2}(\mu_3 + \mu_4) = 0.$$

The linear combination to the left of the equal sign is a contrast. The alternative hypothesis is

$$H_a : \frac{1}{2}(\mu_1 + \mu_2) - \frac{1}{2}(\mu_3 + \mu_4) \neq 0,$$

and a confidence interval is of the form

$$L \leq \frac{1}{2}(\mu_1 + \mu_2) - \frac{1}{2}(\mu_3 + \mu_4) \leq U.$$

If zero is not between L and U, then the null hypothesis is rejected, and we would conclude that color is a significant factor.

An important analytical result involving contrasts is sometimes not completely understood. It is that rejection of the overall null hypothesis (1.1) using an F-test at level of significance α implies that at least one contrast is significant at familywise level of significance α. This may be misinterpreted as implying that the significant contrast(s) is a pairwise difference. In fact, for some data sets the significant contrast(s) will involve more than two means.

In this subsection we discuss using the single-step MCMs for a family of contrasts, not all of which are pairwise. Among the MCMs that assume equal population variances, the choice of the most powerful depends on the number of contrasts in the family and on the number of means in the contrasts. Among the MCMs appropriate when not all population variances are equal, the choice of the most powerful depends on the number of means in the contrasts and on sample sizes.

Scheffe Test (Contrasts). The Scheffe test is also known both as Scheffe's fully significant difference (FSD) test and as Scheffe's globally significant difference (GSD) test. It is exact in that, for the (infinite) family involving all contrasts of k means, the FWER is exactly α. It is also optimal for this family in that, among all simultaneous confidence procedures with familywise confidence coefficient at least $1 - \alpha$, it gives the shortest intervals on average. The Scheffe test should be used when comparisons are selected after looking at the data and they include contrasts, not all of which are pairwise. It should also be considered when a large number of contrasts, not all of

which are pairwise, is specified prior to collecting the data. When the family only involves contrasts of at most three means, the extension of the Tukey test, discussed below, may be more powerful.

Given an FWER of α, the confidence interval for the contrast

$$\sum_{i=1}^{k} c_i \mu_i \tag{2.7}$$

is calculated using the formula

$$\sum_{i=1}^{k} c_i \bar{y}_i \pm \sqrt{(k-1)F_{\alpha,k-1,\nu}} \sqrt{s^2 \sum_{i=1}^{k} \frac{{c_i}^2}{n_i}}, \tag{2.8}$$

where the quantile $F_{\alpha,k-1,\nu}$ is from the F distribution with parameters $k-1$ and ν. The margin of error in (2.8) does not depend on the number of contrasts but does depend on the number of means in the contrast.

Tukey and Tukey–Kramer Tests (Contrasts). The Tukey and Tukey–Kramer tests presented in section 2.2 can be generalized to the family of all contrasts of k means. For the Tukey test, the confidence interval for the contrast (2.7) is calculated using the formula

$$\sum_{i=1}^{k} c_i \bar{y}_i \pm q_{\alpha,\nu,k} \sqrt{\frac{1}{n} s^2} \sum_{i=1}^{k} \frac{|c_i|}{2}. \tag{2.9}$$

For the Tukey–Kramer test, the confidence interval for the contrast is calculated using the formula

$$\sum_{i=1}^{k} c_i \bar{y}_i \pm q_{\alpha,\nu,k} \sqrt{\frac{1}{2} s^2} \frac{\sum_{i=1}^{k} \sum_{j=1}^{k} {c_i}^+ {c_j}^- \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}}{\frac{1}{2} \sum_{i=1}^{k} |c_i|}, \tag{2.10}$$

where ${c_i}^+ = \max(c_i, 0)$ and ${c_j}^- = -\min(c_j, 0)$. The quantile $q_{\alpha,\nu,k}$, is from the Studentized range distribution with parameters ν and k. For both intervals the margin of error is not influenced by the number of contrasts. The confidence intervals will be shorter than Scheffe's intervals for contrasts involving two means and may be shorter for contrasts involving three means.

Bonferroni and Dunn–Šidák Test (Contrasts). Both tests are presented in section 2.2. They are useful for a confirmatory experiment where a family of contrasts is specified prior to collecting the data. Given an FWER of α, the Bonferroni confidence interval for the contrast (2.7) is calculated using the formula

$$\sum_{i=1}^{k} c_i \bar{y}_i \pm t_{\alpha',\nu} \sqrt{s^2 \sum_{i=1}^{k} \frac{{c_i}^2}{n_i}}, \tag{2.11}$$

where $\alpha' = \frac{1}{2}(\alpha/c)$, and the Dunn–Šidák confidence interval for the contrast (2.7) is also calculated using formula (2.11), where

$$\alpha' = \frac{1}{2}\left(1 - (1-\alpha)^{1/c}\right).$$

Table 2 *Selected multiple comparisons using the Scheffe procedure.*

Interval	Significant difference
$-2.102 \le \mu_1 - \mu_2 \le 0.913$	No
$-0.482 \le -\mu_3 + \mu_4 \le 2.533$	No
$0.216 \le -\mu_3 + \mu_5 \le 3.231$	Yes
$-0.809 \le -\mu_4 + \mu_5 \le 2.206$	No
$1.374 \le -\frac{1}{2}\mu_1 - \frac{1}{2}\mu_2 + \mu_3 \le 3.985$	Yes
$2.400 \le -\frac{1}{2}\mu_1 - \frac{1}{2}\mu_2 + \mu_4 \le 5.011$	Yes
$-0.095 \le -\frac{1}{2}\mu_3 - \frac{1}{2}\mu_4 + \mu_5 \le 2.516$	No
$2.663 \le -\mu_1 + \frac{1}{3}\mu_3 + \frac{1}{3}\mu_4 + \frac{1}{3}\mu_5 \le 5.124$	Yes
$2.068 \le -\mu_2 + \frac{1}{3}\mu_3 + \frac{1}{3}\mu_4 + \frac{1}{3}\mu_5 \le 4.530$	Yes
$2.623 \le -\frac{1}{2}\mu_1 - \frac{1}{2}\mu_2 + \frac{1}{3}\mu_3 + \frac{1}{3}\mu_4 + \frac{1}{3}\mu_5 \le 4.569$	Yes
The familywise confidence level is 99%.	
The familywise level of significance is 1%.	

For both formulas, c is the number of contrasts in the family, and the quantile $t_{\alpha',\nu}$ is from Student's t probability distribution with parameter ν. The margin of error in both cases is influenced by the number of specified contrasts. Thus, for small families, the confidence intervals may be shorter than for either the Scheffe test or the Tukey test. The Dunn–Šidák confidence intervals will always be slightly shorter than those from the Bonferroni test. Whenever the Dunn–Šidák test is available, it should be used in place of the Bonferroni test.

Example 2 (multiple comparisons with contrasts using single-step MCMs). Five samples of size 10 were simulated using *Mathematica*. The means of the populations sampled were 5, 6, 8, 9, and 10, respectively, and the variance was 1 for each population. Then each of the Scheffe, the Tukey, the Bonferroni, and the Dunn–Šidák single-step MCMs was run to test for significance of the contrasts in a family of contrasts at the 1% familywise level of significance. The Scheffe test results are reproduced in Table 2. Complete input and output for all of the MCMs in Example 2 are available at the web site given earlier.

The significant differences identified by the Tukey test are identical to those identified by the Scheffe test. The Bonferroni and Dunn–Šidák tests both identified one additional significant difference (for the contrast, $-\frac{1}{2}\mu_3 - \frac{1}{2}\mu_4 + \mu_5$). Table 3 contains the lengths of the confidence intervals in this example. Because the Dunn–Šidák test shows the most differences and has the shortest intervals, it is the most powerful for this problem. Notice that the Bonferroni intervals are only marginally longer than the Dunn–Šidák intervals. The Bonferroni test is more popular with textbook authors, because it is only marginally less powerful and the calculations are easier. Notice that Tukey intervals are shortest for contrasts involving two means. For contrasts involving three means, the Scheffe intervals are slightly shorter than the Tukey intervals. In a different problem, this could be reversed. For contrasts involving more than three means, Scheffe intervals are shorter than Tukey intervals. The Tukey intervals do not necessarily have constant length, as they happen to in this example.

Procedures for Unequal Variances. The GH, C, and T3 procedures discussed in section 2.1 are useful for families of contrasts involving two or three means. For families including contrasts involving more means, a procedure due to Brown and Forsythe [3] is often more powerful. It is an extension of the Scheffe procedure. Like the Scheffe test, it is very conservative for pairwise comparisons.

Table 3 *Length of confidence intervals for each of four single-step MCMs.*

Contrast	Scheffe	Tukey	Bonferroni	Dunn–Šidák
2 means	3.015	2.688	2.734	2.733
3 means	2.611	2.688	2.368	2.367
4 means	2.462	2.688	2.232	2.231
5 means	1.946	2.688	1.765	1.764

2.4. Many-to-One Comparisons. The common form of many-to-one comparisons arises when the means for different experimental treatments are compared to the mean for a control treatment that is designated prior to data collection. The family contains $k-1$ pairwise comparisons, where the control mean is in each comparison. Often, the goal is to identify the experimental treatments with significantly "better" outcome than the control. This requires a one-sided hypothesis test. If the goal is to identify treatments with either significantly "better" or significantly "worse" outcomes, then a two-sided hypothesis test is required.

A second application of many-to-one comparisons arises when the means for different experimental treatments are compared to the mean for the (unknown) "best" treatment. The MCM for this approach is due in large part to Hsu [17]. When the problem is to select the best treatment from several being tested, a natural solution is to choose the treatment with the largest sample mean. But this solution does not take into account the possibility that a "nonbest" treatment could have the largest sample mean. The Hsu procedure will identify treatments that are significantly different from the best treatment. These would be classified as nonbest treatments. Treatments that are not significantly different from the best treatment may be thought of as close to best. A computer program developed by Gupta and Hsu for this MCM is called RSMCB (rank and selection and multiple comparisons with the best) and is included in at least one statistical software package.

Dunnett Test. The Dunnett test [6] is an exact test for the family of many-to-one comparisons when the control is designated prior to data collection. Because the control group is of such importance, it should have a larger sample size than the other groups. The following rule of thumb tends to optimize the procedure. If each of the $k-1$ experimental groups has n values, then the control group should have approximately $n\sqrt{k-1}$ values. If the experimental groups have different sample sizes, the mean sample size is multiplied by $\sqrt{k-1}$ to decide on a sample size for the control group. The two-sided test is of the form

$$\begin{aligned} H_0 &: \mu_i = \mu_c, \\ H_a &: \mu_i \neq \mu_c, \end{aligned} \tag{2.12}$$

where μ_i is the mean of an experimental population and μ_c is the mean of the control population. A significant difference is shown when the confidence interval for $\mu_i - \mu_c$ does not contain zero. For a lower one-sided test with corresponding alternative hypothesis $H_a : \mu_i - \mu_c < 0$, a significant difference is shown when the upper one-sided confidence limit is less than zero. For an upper one-sided test with corresponding alternative hypothesis $H_a : \mu_i - \mu_c > 0$, a significant difference is shown when the lower one-sided confidence limit is greater than zero.

Table 4 *Multiple pairwise comparisons with a control using the Dunnett procedure.*

i	c	Confidence intervals	Significant difference
1	4	$\mu_1 - \mu_4 \leq -0.256$	Yes
2	4	$\mu_2 - \mu_4 \leq 1.178$	No
3	4	$\mu_3 - \mu_4 \leq 0.465$	No
The familywise confidence coefficient is 95%.			
The familywise level of significance is 5%.			

The two-sided confidence interval for $\mu_i - \mu_c$ is calculated using the formula

$$\bar{y}_i - \bar{y}_c \pm |T|_{\alpha,k-1,\nu,[\rho_{i,j}]} \sqrt{s^2 \left(\frac{1}{n_i} + \frac{1}{n_c} \right)}; \tag{2.13}$$

the upper one-sided confidence limit is calculated using

$$\bar{y}_i - \bar{y}_c + T_{\alpha,k-1,\nu,[\rho_{i,j}]} \sqrt{s^2 \left(\frac{1}{n_i} + \frac{1}{n_c} \right)}; \tag{2.14}$$

and the lower one-sided confidence limit is calculated using

$$\bar{y}_i - \bar{y}_c - T_{\alpha,k-1,\nu,[\rho_{i,j}]} \sqrt{s^2 \left(\frac{1}{n_i} + \frac{1}{n_c} \right)}. \tag{2.15}$$

Because the comparisons in the family all contain the control mean, the correlation coefficient for the ith mean versus the control and the jth mean versus the control, denoted by $\rho_{i,j}$, must be considered for all $i \neq j$. When the sample sizes are all equal, $\rho_{i,j} = 0.5$ for all $i \neq j$, and tables are available for the quantiles $|T|_{\alpha,k-1,\nu,[\rho_{i,j}]}$ and $T_{\alpha,k-1,\nu,[\rho_{i,j}]}$. When the sample size for the control group is different, but the experimental groups all have the same sample size, then the correlation coefficients are constant. The general formula is

$$\rho_{i,j} = \sqrt{\frac{n_i n_j}{(n_i + n_c)(n_j + n_c)}}.$$

In this case, calculation of the quantiles $|T|_{\alpha,k-1,\nu,[\rho_{i,j}]}$ and $T_{\alpha,k-1,\nu,[\rho_{i,j}]}$ requires repeated evaluation of a double integral of a product involving $k-1$ terms.

Example 3 (Dunnett test). In Example 1, four samples of sizes 10, 10, 10, and 20, respectively, were simulated. The means of the populations sampled were 5, 6, 6, and 6, respectively, and the variance was 1 for each population. Assuming that the fourth sample is from the control population, a one-sided Dunnett test was done for the family of three pairwise comparisons to determine which of the other populations has a mean lower than the control mean. The results are reproduced in Table 4. Complete input and output for Example 3 are available at the web site. We conclude at the 5% familywise level of significance that $\mu_1 < \mu_4$.

3. Step-Down Procedures. Two well-known step-down procedures are the Newman–Keuls (sometimes called the Student–Newman–Keuls) test and the Duncan multiple range test (the most liberal of all MCMs). As they were originally presented and commonly used, these tests do not protect the FWER. The four procedures presented in this section do protect the specified FWER.

There is good reason to consider multiple-step procedures, because they are often more powerful than their single-step counterparts. However, there are also some drawbacks that may severely limit their usefulness. First, they can usually only be used for hypothesis testing. For most of these procedures it is not known how to obtain corresponding simultaneous confidence intervals. Second, they can only be used for finite families, and the inferences usually cannot be extended to a larger family than was initially specified [16]. Third, the impact of directional decisions is not known. Recall from Example 1 that for single-step MCMs, the chance of making errors in directional decisions is included in the specified FWER along with the chance of making errors in decisions concerning differences. For most step-down procedures, it is not known whether the chance of making errors in directional decisions is similarly included in the specified FWER. There is the possibility that directional decisions would increase the actual FWER.

The development of the first two procedures, called the multiple F step-down test and the multiple Q step-down test, is attributed to various authors, and the procedures may be called by different names depending on the reference consulted. Initially, Ryan [19] proposed a modification to the Newman–Keuls procedure to protect the FWER. The proposal was modified first by Einot and Gabriel [10] and then by Welsch [21]. In some implementations the multiple F step-down test is named the REGWF procedure, and the multiple Q step-down test is named the REGWQ procedure.

The procedures begin by testing all k means. If significance is found, then all subsets consisting of $k-1$ means are tested. In general, if a set consisting of p means is significant, then all of its subsets of size $p-1$ are tested. If a set of p means is not significant, then none of its subsets is tested. All such subsets are assumed to be nonsignificant. The procedure ultimately identifies the maximal nonsignificant subsets of the k means, where a set is *maximal nonsignificant* if it is not significant and not a proper subset of another nonsignificant set. All of the means that occupy the same maximal nonsignificant subset are not significantly different, and means that do not occupy the same maximal nonsignificant subset are significantly different. Significance or nonsignificance of any subset of the means can be ascertained from the lists of maximal nonsignificant subsets.

3.1. Multiple *F* Step-Down Test. To test for significance in a set P consisting of p means, the test statistic is

$$F_P = \frac{1}{(p-1)s^2}\left(\sum_{i\in P} n_i \bar{y}_i^2 - \frac{\left(\sum_{i\in P} \bar{y}_i\right)^2}{\sum_{i\in P} n_i}\right).$$

The test is significant at familywise level of significance α if $F_P \geq F_{\alpha_p,p-1,\nu}$, where the quantile $F_{\alpha_p,p-1,\nu}$ is calculated from the F distribution with parameters $p-1$ and ν and where

$$\alpha_p = 1-(1-\alpha)^{p/k}, \quad 2 \leq p \leq k-2, \quad \alpha_{k-1} = \alpha_k = \alpha. \tag{3.1}$$

3.2. Multiple *Q* Step-Down Test. To test for significance in a set P consisting of p means when sample sizes are equal, the test statistic is

$$Q_P = \frac{\max \bar{y} - \min \bar{y}}{s/\sqrt{n}},$$

where $\max \bar{y}$ and $\min \bar{y}$ are, respectively, the largest and smallest means in P. The test is significant at familywise level of significance α if $Q_P \geq q_{\alpha_p,\nu,p}$, where the quantile

Table 5 *Multiple pairwise comparisons using the PeritzQ procedure.*

i	j	$\lvert\bar{y}_i - \bar{y}_j\rvert$	Significant difference
3	7	202	Yes
1	7	265	Yes
1	6	222	Yes
The maximal nonsignificant subsets are			
{{1,1,1,1,1,0,0},{0,1,1,1,1,1,0},{0,1,0,1,1,1,1}}			

$q_{\alpha_p,\nu,p}$ is calculated from the Studentized range distribution with parameters p and ν and where α_p is defined in (3.1). In case the sample sizes are not all equal, the test statistic is

$$Q_P = \max\left\{\frac{|\bar{y}_i - \bar{y}_j|}{\sqrt{\frac{1}{2}s^2\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}}\right\},$$

where the maximum is taken for all possible pairs of means in P.

The other two procedures presented here are refinements of the multiple F step-down test and the multiple Q step-down test. An algorithm developed by Begun and Gabriel [2] is used to incorporate a method due to Peritz that looks for additional differences. Thus the PeritzF test is at least as powerful as the multiple F step-down test, and the PeritzQ test is at least as powerful as the multiple Q step-down test. The Peritz tests are not as commonly available in statistical software packages.

Example 4 (step-down procedures). Data taken from an article by Duncan [4] is used to illustrate the step-down procedures. There were seven random samples of sizes 3, 2, 5, 5, 3, 2, and 3 with respective sample means of 680, 734, 743, 851, 873, 902, and 945. The results of the PeritzQ test are reproduced in Table 5. The commands and output for all four step-down tests in Example 4 are available on the web site.

In addition to indicating the pairs of means that differ significantly, the output also indicates the maximal nonsignificant subsets. For example, the list {0,1,0,1,1,1,1} describes a maximal nonsignificant subset of five means. It indicates that a significant difference cannot be shown among μ_2, μ_4, μ_5, μ_6, and μ_7. Significant differences at the 5% familywise level of significance are shown between μ_1 and μ_6, between μ_1 and μ_7, and between μ_3 and μ_7. None of these pairs occurs together in the maximal nonsignificant subsets. The results from the multiple Q test are identical to the PeritzQ results. Both the PeritzF test and the multiple F test identify two significant differences (between μ_3 and μ_7 and between μ_1 and μ_7). Simulation studies suggest that the choice of the most powerful test between the PeritzQ test and the PeritzF test or between the multiple Q step-down test and the multiple F step-down test will depend on the family and the data.

4. Choosing and Using Procedures. Given the variety of statistical software packages available to perform multiple comparisons, a user need not be concerned about doing the calculations. But the user should have the knowledge necessary to correctly choose the appropriate MCMs and correctly interpret the computer output. When choosing among the MCMs that protect the specified FWER, a user should be aware of several considerations.

4.1. The Family. Depending on the type of comparisons in the family, one method may be more appropriate to use than others. For example, Tukey's test is most useful for the family of all pairwise comparisons; Scheffe's test is most useful when the family consists of a great many (or infinitely many) contrasts, not all of which are pairwise; the GT2, Dunn–Šidák, and Bonferroni tests are most useful when smaller numbers of pairwise comparisons and/or other contrasts are prespecified; the Dunnett test is used for the family of comparisons of $k-1$ experimental means to a control mean; and the GH, C, T3, and Brown–Forsythe procedures are used when the population variances are not all equal.

4.2. Exploratory and Confirmatory. The comparisons to be made may be planned before looking at the data, or they may be selected after looking at the data. Terms such as *data-snooping* and *post hoc selection* have been used to describe the latter practice. When the comparisons to be made are specified before looking at the data, the selection of the appropriate MCM should be based on the family. For example, if a set of contrasts, which may include selected pairwise comparisons, is specified, then the GT2 (Dunn–Šidák or Bonferroni) and Scheffe tests should be tried. On the other hand, if prior to viewing the data the intention is to investigate all pairwise differences, then the Tukey or Tukey–Kramer test and perhaps the stepwise procedures should be tried. The user must keep in mind that directional decisions and confidence intervals may not be possible with the stepwise procedures.

If a specific set of comparisons is developed after looking at the data, then a procedure for the family from which the comparisons are selected should be chosen. For example, if the selected comparisons are pairwise, then the family would be all pairwise comparisons or, if appropriate, the family of comparisons of several experimental means to a control mean. If the set of comparisons includes more general contrasts, then the family would include all possible contrasts of a set of k means (an infinite family).

In case some of the comparisons are prespecified and some are suggested by the data, it is appropriate to perform the prespecified comparisons using FWER α_1 (say) and the post hoc comparisons using FWER α_2 (say) on the family from which these comparisons are selected. Then the overall FWER would be at most $\alpha_1 + \alpha_2$.

4.3. Error Rate. Some methods are exact, so that the actual FWER equals the value specified by the researcher. These include the Tukey, Scheffe, and Dunnett tests. Others are based on approximations and are conservative in that the actual FWER is less than the value specified. These include the Tukey–Kramer, GT2, Bonferroni, Dunn–Šidák, C, T3, and Brown–Forsythe tests. Still others, like the GH test, may at times be liberal. That is, depending on the data, the actual FWER may be greater than the value specified. Exact methods or those that are minimally conservative are usually the best choice.

4.4. Assumptions. An MCM may or may not be robust in the presence of a violation of one or more of its underlying assumptions. No MCM is robust to violations of randomness or independence. Most MCMs seem to be robust to moderate departures from normality of the underlying populations in that the actual FWER will only be slightly higher than that specified. However, this increase in actual error rate may be magnified as the number of comparisons increases. For balanced designs the appropriate MCMs appear to be robust to the effect of unequal population variances. This, however, may not be the case for unbalanced designs. For example, under some configurations of sample sizes and population variances, the Tukey–Kramer test may

have a severely inflated actual FWER. There are alternative procedures to use when the variances are not all equal.

4.5. Power. After all of the pertinent considerations, it may happen that two (or more) MCMs are appropriate for a family, but it may not be clear if one is more powerful than the other. If both MCMs are run at the same FWER and show the same differences, then one or the other may be used. On the other hand, if one shows more differences, then using the strategy of always choosing the more powerful has the associated risk of an inflated FWER. Choosing some differences from one procedure and other differences from the other will likely double the FWER. Finally, choosing to report the results of the MCM that shows a difference of particular importance will also inflate the FWER. This is especially not recommended if the study is confirmatory. For a confirmatory study where one difference is particularly important, consulting the literature or using other criteria to choose an MCM that appears to give the best chance to show this difference is advised [20, Chapter 3].

4.6. Simulation. Much of what is known about the commonly used MCMs has been learned through simulation. For example, data sets may be randomly generated based on different configurations of the underlying population means. In a minimum range configuration, the first half of the means are equal and the second half of the means are equal but different from the first half. In a maximum range configuration, the first mean is lowest, the last mean is highest, and the middle means are equal and half-way between the lowest and highest means. In an equally spaced configuration, each succeeding mean is a fixed increment higher than its preceding mean. Using *Mathematica*, after a few hours, any interested user can begin research into the behavior of these MCMs for a variety of population distributions and a variety of configurations of the population means, population variances, and sample sizes. As an illustration, we have included on the web site (as Example 5) an investigation of robustness for the Tukey–Kramer test.

REFERENCES

[1] M. L. Abell, J. P. Braselton, and J. A. Rafter, *Statistics with Mathematica*, Academic Press, Boston, 1999.

[2] J. Begun and K. R. Gabriel, *Closure of the Newman-Keuls multiple comparisons procedure*, J. Amer. Statist. Assoc., 76 (1981), pp. 241–245.

[3] M. B. Brown and A. B. Forsythe, *The ANOVA and multiple comparisons for data with heterogeneous variances*, Biometrics, 30 (1974), pp. 719–724.

[4] D. B. Duncan, *Multiple range tests for correlated and heteroscedastic means*, Biometrics, 13 (1957), pp. 164–176.

[5] O. J. Dunn, *Estimation of the means of dependent variables*, Ann. Math. Statist., 29 (1958), pp. 1095–1111.

[6] C. W. Dunnett, *A multiple comparison procedure for comparing several treatments with a control*, J. Amer. Statist. Assoc., 50 (1955), pp. 1096–1121.

[7] C. W. Dunnett, *Pairwise multiple comparisons in the homogeneous variance, unequal sample size case*, J. Amer. Statist. Assoc., 75 (1980), pp. 789–795.

[8] C. W. Dunnett, *Pairwise multiple comparisons in the unequal variance case*, J. Amer. Statist. Assoc., 75 (1980), pp. 796–800.

[9] C. W. Dunnett and A. C. Tamhane, *A step-up multiple test procedure*, J. Amer. Statist. Assoc., 87 (1992), pp. 162–170.

[10] I. Einot and K. R. Gabriel, *A study of the powers of several methods of multiple comparisons*, J. Amer. Statist. Assoc., 70 (1975), pp. 574–583.

[11] R. A. Fisher, *The Design of Experiments*, Oliver & Boyd, Edinburgh, 1935.

[12] K. R. Gabriel, *On the relation between union-intersection and likelihood ratio tests*, in Essays in Probability and Statistics, R. C. Bose et al., eds., University of North Carolina Press, Chapel Hill, 1970, pp. 251–266.

[13] P. A. Games and J. F. Howell, *Pairwise multiple comparison procedures with unequal N's and/or variances: A Monte Carlo study*, J. Educ. Statist., 1 (1976), pp. 113–125.
[14] A. J. Hayter, *A proof of the conjecture that the Tukey-Kramer multiple comparison procedure is conservative*, Ann. Statist., 12 (1984), pp. 61–75.
[15] Y. Hochberg, *A sharper Bonferroni procedure for multiple tests of significance*, Biometrika, 75 (1988), pp. 800–802.
[16] Y. Hochberg and A. C. Tamhane, *Multiple Comparison Procedures*, Wiley, New York, 1987.
[17] J. C. Hsu, *Multiple Comparisons: Theory and Methods*, Chapman & Hall, UK, 1996.
[18] J. Neter, M. H. Kutner, C. J. Nachtshein, and W. Wasserman, *Applied Linear Statistical Models*, 4th ed., Irwin, Chicago, 1996.
[19] T. A. Ryan, *Significance tests for multiple comparison of proportions, variances and order statistics*, Psychol. Bull., 57 (1960), pp. 318–328
[20] L. E. Toothaker, *Multiple Comparisons for Researchers*, Sage, Newbury Park, CA, 1991.
[21] R. E. Welsch, *Stepwise multiple comparison procedures*, J. Amer. Statist. Assoc., 72 (1977), pp. 566–575.