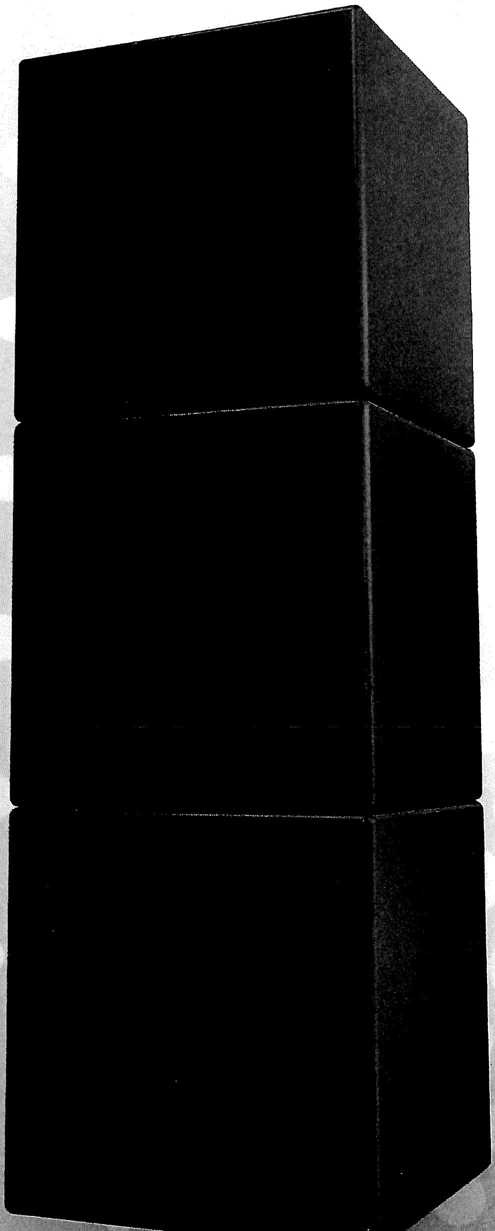
Experimental Design

4TH EDITION

2013

Procedures for the Behavioral Sciences



ROGER E.
KIRK

CHAPTER 5

Multiple Comparison Tests

5.1 Introduction to Multiple Comparison Tests

The most common use of analysis of variance is in testing the hypothesis that $p \geq 3$ population means are equal. If the omnibus hypothesis of equality of means is rejected, a researcher is still faced with the problem of deciding which of the means are not equal. Thus, an omnibus *F* test is merely one step in analyzing a set of data. A significant *F* test indicates that something has happened in an experiment that has a small probability of happening by chance. In this chapter, I describe a variety of procedures for pinpointing what has happened. Specifically, I examine a number of test statistics for deciding which population means are not equal. But first, I need to introduce some important concepts.

Contrasts Among Means

A contrast or comparison among means is a difference among the means, with appropriate algebraic signs. I use the symbols ψ_i and $\hat{\psi}_i$ to denote, respectively, the *i*th contrast among population means and a sample estimate of the *i*th contrast. For example, $\psi_i = \mu_j - \mu_{j'}$ is a contrast for population means μ_j and $\mu_{j'}$; $\hat{\psi}_i = \overline{Y}_{\cdot j} - \overline{Y}_{\cdot j'}$ is a sample estimator of the population contrast. If an experiment contains $p = 3$ treatment levels, contrasts involving two and three means may be of interest—for example,

$$\begin{array}{lll} & \hat{\psi}_1 = \overline{Y}_{\cdot 1} - \overline{Y}_{\cdot 2} & \hat{\psi}_4 = \dfrac{\overline{Y}_{\cdot 1} + \overline{Y}_{\cdot 2}}{2} - \overline{Y}_{\cdot 3} \\ (5.1\text{-}1) & \hat{\psi}_2 = \overline{Y}_{\cdot 1} - \overline{Y}_{\cdot 3} & \hat{\psi}_5 = \dfrac{\overline{Y}_{\cdot 1} + \overline{Y}_{\cdot 3}}{2} - \overline{Y}_{\cdot 2} \\ & \hat{\psi}_3 = \overline{Y}_{\cdot 2} - \overline{Y}_{\cdot 3} & \hat{\psi}_6 = \dfrac{\overline{Y}_{\cdot 2} + \overline{Y}_{\cdot 3}}{2} - \overline{Y}_{\cdot 1} \end{array}$$

The contrasts on the right involve the average of two means versus a third mean. Such contrasts could be used, for example, to compare the average of two experimental groups with a control group.

More formally, a **contrast** or **comparison** among means is a linear combination of means that have known weights or coefficients. The coefficients are denoted by c_j and satisfy two conditions: (1) at least one coefficient is not equal to zero ($c_j \neq 0$ for some j), and (2) the coefficients sum to zero ($\sum_{j=1}^{p} c_j = 0$). The contrasts

$$\psi_i = c_1\mu_1 + c_2\mu_2 + \cdots + c_p\mu_p$$

and

$$\hat{\psi}_i = c_1\overline{Y}_{\cdot 1} + c_2\overline{Y}_{\cdot 2} + \cdots + c_p\overline{Y}_{\cdot p}$$

are, respectively, population and sample contrasts if $c_j \neq 0$ for some j and $\sum_{j=1}^{p} c_j = 0$. The contrasts in equations (5.1-1) can be expressed as linear combinations of sample means by the appropriate choice of coefficients:

$$\begin{array}{cccccccl}
\hat{\psi}_i & = & c_1\overline{Y}_{\cdot 1} & + & c_2\overline{Y}_{\cdot 2} & + & c_3\overline{Y}_{\cdot 3} & \\
\hline
\hat{\psi}_1 & = & 1\overline{Y}_{\cdot 1} & + & (-1)\overline{Y}_{\cdot 2} & + & 0\overline{Y}_{\cdot 3} & = \overline{Y}_{\cdot 1} - \overline{Y}_{\cdot 2} \\
\hat{\psi}_2 & = & 1\overline{Y}_{\cdot 1} & + & 0\overline{Y}_{\cdot 2} & + & (-1)\overline{Y}_{\cdot 3} & = \overline{Y}_{\cdot 1} - \overline{Y}_{\cdot 3} \\
\hat{\psi}_3 & = & 0\overline{Y}_{\cdot 1} & + & 1\overline{Y}_{\cdot 2} & + & (-1)\overline{Y}_{\cdot 3} & = \overline{Y}_{\cdot 2} - \overline{Y}_{\cdot 3} \\
\hat{\psi}_4 & = & \frac{1}{2}\overline{Y}_{\cdot 1} & + & \frac{1}{2}\overline{Y}_{\cdot 2} & + & (-1)\overline{Y}_{\cdot 3} & = \dfrac{\overline{Y}_{\cdot 1} + \overline{Y}_{\cdot 2}}{2} - \overline{Y}_{\cdot 3} \\
\hat{\psi}_5 & = & \frac{1}{2}\overline{Y}_{\cdot 1} & + & (-1)\overline{Y}_{\cdot 2} & + & \frac{1}{2}\overline{Y}_{\cdot 3} & = \dfrac{\overline{Y}_{\cdot 1} + \overline{Y}_{\cdot 3}}{2} - \overline{Y}_{\cdot 2} \\
\hat{\psi}_6 & = & (-1)\overline{Y}_{\cdot 1} & + & \frac{1}{2}\overline{Y}_{\cdot 2} & + & \frac{1}{2}\overline{Y}_{\cdot 3} & = \dfrac{\overline{Y}_{\cdot 2} + \overline{Y}_{\cdot 3}}{2} - \overline{Y}_{\cdot 1}
\end{array} \tag{5.1-2}$$

Notice that for each contrast, $c_j \neq 0$ for some j and $\sum_{j=1}^{p} c_j = 0$. For convenience in comparing the magnitudes of different contrasts, the coefficients of each contrast can be chosen so that the sum of their absolute values is equal to 2; that is,

$$\sum_{j=1}^{p} |\, c_j \,| = 2$$

where $|\, c_j \,|$ indicates the absolute value of c_j and is equal to the positive member of c_j and $-c_j$. All six of the preceding contrasts satisfy $\sum_{j=1}^{p} |\, c_j = 2 |$. For example, the sum of the absolute values of the coefficients for $\hat{\psi}_1$ and $\hat{\psi}_4$ is, respectively,

$$|1| + |-1| + |\,0\,| = 1 + 1 + 0 = 2$$

$$\left|\tfrac{1}{2}\right| + \left|\tfrac{1}{2}\right| + |-1| = \tfrac{1}{2} + \tfrac{1}{2} + 1 = 2$$

Pairwise and Nonpairwise Comparisons

When all of the coefficients of a contrast except two are equal to zero, the contrast is called a **pairwise comparison;** otherwise, the contrast is a **nonpairwise comparison.** The number of pairwise comparisons that exist for p means is equal to $p(p - 1)/2$. For example, contrasts $\hat{\psi}_1$, $\hat{\psi}_2$, and $\hat{\psi}_3$ in equations (5.1-2) exhaust the $3(3 - 1)/2 = 3$ pairwise comparisons among three means. The situation is quite different for nonpairwise comparisons—the number is infinite. Consider the following examples in which an average of two means is compared with a third mean:

$$\hat{\psi}_4 = \tfrac{1}{2}\overline{Y}_{\cdot 1} + \tfrac{1}{2}\overline{Y}_{\cdot 2} + (-1)\overline{Y}_{\cdot 3} = \frac{1\overline{Y}_{\cdot 1} + 1\overline{Y}_{\cdot 2}}{2} - \overline{Y}_{\cdot 3}$$

$$\hat{\psi}_7 = \tfrac{1}{3}\overline{Y}_{\cdot 1} + \tfrac{2}{3}\overline{Y}_{\cdot 2} + (-1)\overline{Y}_{\cdot 3} = \frac{1\overline{Y}_{\cdot 1} + 2\overline{Y}_{\cdot 2}}{3} - \overline{Y}_{\cdot 3}$$

$$\hat{\psi}_8 = \tfrac{1}{4}\overline{Y}_{\cdot 1} + \tfrac{3}{4}\overline{Y}_{\cdot 2} + (-1)\overline{Y}_{\cdot 3} = \frac{1\overline{Y}_{\cdot 1} + 3\overline{Y}_{\cdot 2}}{4} - \overline{Y}_{\cdot 3}$$

$$\hat{\psi}_9 = \tfrac{1}{5}\overline{Y}_{\cdot 1} + \tfrac{4}{5}\overline{Y}_{\cdot 2} + (-1)\overline{Y}_{\cdot 3} = \frac{1\overline{Y}_{\cdot 1} + 4\overline{Y}_{\cdot 2}}{5} - \overline{Y}_{\cdot 3}$$

The coefficients $\frac{1}{3}$ and $\frac{2}{3}$ in contrast 7, for example, indicate that $\overline{Y}_{\cdot 2}$ is weighted twice as much as $\overline{Y}_{\cdot 1}$ when the means are averaged. The pattern of coefficients $\frac{1}{2}, \frac{1}{2}, -1$; $\frac{1}{3}, \frac{2}{3}, -1$; $\frac{1}{4}, \frac{3}{4}, -1$; and $\frac{1}{5}, \frac{4}{5}, -1$ can be continued indefinitely. Hence, an infinite number of nonpairwise contrasts can be constructed. Notice that the coefficients of these nonpairwise contrasts are selected so that they satisfy the optional requirement that $\sum_{j=1}^{p}\left| c_j = 2 \right|$.

Orthogonal Contrasts

An infinite number of contrasts can be constructed for $p \geq 3$ means. Each of these contrasts can be expressed as a linear combination of $p - 1$ contrasts. For example, the contrast $\hat{\psi}_2 = \overline{Y}_{\cdot 1} - \overline{Y}_{\cdot 3}$ in equations (5.1-2) is equal to $\frac{1}{2}\hat{\psi}_1 + \hat{\psi}_4$:

$$\hat{\psi}_2 = \overbrace{\left[\tfrac{1}{2}(\overline{Y}_{\cdot 1} - \overline{Y}_{\cdot 2})\right]}^{\frac{1}{2}\hat{\psi}_1} + \overbrace{\left[\tfrac{1}{2}\overline{Y}_{\cdot 1} + \tfrac{1}{2}\overline{Y}_{\cdot 2} - \overline{Y}_{\cdot 3}\right]}^{\hat{\psi}_4} = \overline{Y}_{\cdot 1} - \overline{Y}_{\cdot 3}$$

Thus, contrast 2 provides no information that cannot be obtained from contrasts 1 and 4. Similarly, contrast 3, $\hat{\psi}_3 = \overline{Y}_{\cdot 2} - \overline{Y}_{\cdot 3}$, is equal to $-\frac{1}{2}\hat{\psi}_1 + \hat{\psi}_4$:

$$\hat{\psi}_3 = \overbrace{\left[(-\tfrac{1}{2})(\overline{Y}_{\cdot 1} - \overline{Y}_{\cdot 2})\right]}^{-\frac{1}{2}\hat{\psi}_1} + \overbrace{\left[\tfrac{1}{2}\overline{Y}_{\cdot 1} + \tfrac{1}{2}\overline{Y}_{\cdot 2} - \overline{Y}_{\cdot 3}\right]}^{\hat{\psi}_4} = \overline{Y}_{\cdot 2} - \overline{Y}_{\cdot 3}$$

Contrasts 2 and 3 are redundant because they can be expressed as linear combinations of contrasts 1 and 4.

Sometimes a researcher is interested in contrasts that are mutually nonredundant. Such contrasts are called **orthogonal contrasts.** There is a simple rule for determining whether two contrasts are orthogonal. Let $\hat{\psi}_i$ and $\hat{\psi}_{i'}$ denote the ith and i'th contrasts and c_{ij} and $c_{i'j}$ their respective coefficients, where $j = 1, \ldots, p$. The two contrasts are orthogonal if

$$\sum_{j=1}^{p} c_{ij}c_{i'j} = 0$$

for the equal n case or

$$\sum_{j=1}^{p} \frac{c_{ij}c_{i'j}}{n_j} = 0$$

for the unequal n case. Consider the contrasts in set 1:

Set 1

$$\hat{\psi}_1 = 1\overline{Y}_{\cdot 1} + (-1)\overline{Y}_{\cdot 2} + 0\overline{Y}_{\cdot 3}$$

$$\hat{\psi}_4 = \tfrac{1}{2}\overline{Y}_{\cdot 1} + \tfrac{1}{2}\overline{Y}_{\cdot 2} + (-1)\overline{Y}_{\cdot 3}$$

and assume that the n_js are equal. These two contrasts are orthogonal because the sum of the products of their coefficients is zero:

$$\sum_{j=1}^{p} c_{1j}c_{4j} = (1)(\tfrac{1}{2}) + (-1)(\tfrac{1}{2}) + (0)(-1) = 0$$

However, contrasts

$$\hat{\psi}_1 = 1\overline{Y}_{\cdot 1} + (-1)\overline{Y}_{\cdot 2} + 0\overline{Y}_{\cdot 3} \qquad \text{and} \qquad \hat{\psi}_2 = 1\overline{Y}_{\cdot 1} + 0\overline{Y}_{\cdot 2} + (-1)\overline{Y}_{\cdot 3}$$

are not orthogonal because

$$\sum_{j=1}^{p} c_{1j}c_{2j} = (1)(1) + (-1)(0) + (0)(-1) = 1$$

Contrasts $\hat{\psi}_1$ and $\hat{\psi}_4$ are one of the infinite number of sets of orthogonal contrasts among three means. Three other sets of orthogonal contrasts are

Set 2

$$\hat{\psi}_2 = 1\overline{Y}_{\cdot 1} + 0\overline{Y}_{\cdot 2} + (-1)\overline{Y}_{\cdot 3}$$

$$\hat{\psi}_5 = \tfrac{1}{2}\overline{Y}_{\cdot 1} + (-1)\overline{Y}_{\cdot 2} + \tfrac{1}{2}\overline{Y}_{\cdot 3}$$

$$\text{Set 3} \qquad \begin{aligned} \hat{\psi}_3 &= 0\bar{Y}_{.1} + 1\bar{Y}_{.2} + (-1)\bar{Y}_{.3} \\ \hat{\psi}_6 &= (-1)\bar{Y}_{.1} + \tfrac{1}{2}\bar{Y}_{.2} + \tfrac{1}{2}\bar{Y}_{.3} \end{aligned}$$

$$\text{Set 4} \qquad \begin{aligned} \hat{\psi}_7 &= \tfrac{1}{3}\bar{Y}_{.1} + \tfrac{2}{3}\bar{Y}_{.2} + (-1)\bar{Y}_{.3} \\ \hat{\psi}_{10} &= (-1)\bar{Y}_{.1} + \tfrac{4}{5}\bar{Y}_{.2} + \tfrac{1}{5}\bar{Y}_{.3} \end{aligned}$$

because

$$\begin{aligned} \sum_{j=1}^{p} c_{2j}c_{5j} &= (1)(\tfrac{1}{2}) + (0)(-1) + (-1)(\tfrac{1}{2}) = 0 \\ \sum_{j=1}^{p} c_{3j}c_{6j} &= (0)(-1) + (1)(\tfrac{1}{2}) + (-1)(\tfrac{1}{2}) = 0 \\ \sum_{j=1}^{p} c_{7j}c_{10j} &= (\tfrac{1}{3})(-1) + (\tfrac{2}{3})(\tfrac{4}{5}) + (-1)(\tfrac{1}{5}) = 0 \end{aligned}$$

I have now identified four of the infinite number of sets of orthogonal contrasts among three means. Consider the set $\hat{\psi}_1$ and $\hat{\psi}_4$ again. The reader may wonder if it is possible to find another contrast that is orthogonal to $\hat{\psi}_1$ and $\hat{\psi}_4$. The answer is no. The maximum number of orthogonal contrasts in any set is equal to $p - 1$. For my example, that number is $3 - 1 = 2$. To summarize, for $p \geq 3$ means, there are an infinite number of sets of orthogonal contrasts, but each set contains only $p - 1$ orthogonal contrasts.

It can be shown that any orthogonal set of $p - 1$ contrasts provides a basis for constructing all other contrasts that involve p means; that is, all contrasts can be expressed as linear combinations of the contrasts in an orthogonal set. For example, I showed that $\hat{\psi}_1$ and $\hat{\psi}_4$ are orthogonal and that they could be used to construct $\hat{\psi}_2$ and $\hat{\psi}_3$:

$$\hat{\psi}_2 = \tfrac{1}{2}\hat{\psi}_1 + \hat{\psi}_4$$

and

$$\hat{\psi}_3 = (-\tfrac{1}{2})\hat{\psi}_1 + \hat{\psi}_4$$

I show here that contrasts $\hat{\psi}_5, \ldots, \hat{\psi}_{10}$ also can be expressed as linear combinations of $\hat{\psi}_1 = \bar{Y}_{.1} + \bar{Y}_{.2}$ and $\hat{\psi}_4 = \tfrac{1}{2}\bar{Y}_{.1} + \tfrac{1}{2}\bar{Y}_{.2} + (-1)\bar{Y}_{.3}$:

$$\begin{aligned} \hat{\psi}_5 &= \tfrac{3}{4}\hat{\psi}_1 + (-\tfrac{1}{2})\hat{\psi}_4 = \frac{\bar{Y}_{.1} + \bar{Y}_{.3}}{2} - \bar{Y}_{.2} \\ \hat{\psi}_6 &= (-\tfrac{3}{4})\hat{\psi}_1 + (-\tfrac{1}{2})\hat{\psi}_4 = \frac{\bar{Y}_{.2} + \bar{Y}_{.3}}{2} - \bar{Y}_{.1} \\ \hat{\psi}_7 &= (-\tfrac{1}{6})\hat{\psi}_1 + \hat{\psi}_4 = \frac{1\bar{Y}_{.1} + 2\bar{Y}_{.2}}{3} - \bar{Y}_{.3} \end{aligned}$$

$$\hat{\psi}_8 = (-\tfrac{1}{4})\hat{\psi}_1 + \hat{\psi}_4 = \frac{1\overline{Y}_{\cdot 1} + 3\overline{Y}_{\cdot 2}}{4} - \overline{Y}_{\cdot 3}$$

$$\hat{\psi}_9 = (-\tfrac{3}{10})\hat{\psi}_1 + \hat{\psi}_4 = \frac{1\overline{Y}_{\cdot 1} + 4\overline{Y}_{\cdot 2}}{5} - \overline{Y}_{\cdot 3}$$

$$\hat{\psi}_{10} = (-\tfrac{9}{10})\hat{\psi}_1 + (-\tfrac{1}{5})\hat{\psi}_4 = \frac{4\overline{Y}_{\cdot 2} + 1\overline{Y}_{\cdot 3}}{5} - \overline{Y}_{\cdot 1}$$

As I have shown, there are always $p - 1$ nonredundant questions that can be answered from the data in an experiment. However, a researcher may not be interested in all of the $p - 1$ questions. For example, in an experiment with three means, a researcher may want to test the hypothesis that $\mu_1 - \mu_2 = 0$ but not that $\frac{1}{2}\mu_1 + \frac{1}{2}\mu_2 + (-1)\mu_3 = 0$. The second hypothesis, which is orthogonal to the first, may have no meaning in terms of the objectives of the experiment. Also, many interesting research questions involve nonorthogonal contrasts. In an experiment with three treatment levels, each of the three pairwise contrasts among means may be associated with a question that the researcher wants to answer. However, a researcher who tests the three pairwise contrasts needs to understand that the tests involve redundant information. For example, the value of contrast $\hat{\psi}_3 = \overline{Y}_{\cdot 2} - \overline{Y}_{\cdot 3}$ can be obtained from contrasts $\hat{\psi}_1 = \overline{Y}_{\cdot 1} - \overline{Y}_{\cdot 2}$ and $\hat{\psi}_2 = \overline{Y}_{\cdot 1} - \overline{Y}_{\cdot 3}$ as follows:

$$\hat{\psi}_3 = \overbrace{\left(\overline{Y}_{\cdot 1} - \overline{Y}_{\cdot 3}\right)}^{\hat{\psi}_2} - \overbrace{\left(\overline{Y}_{\cdot 1} - \overline{Y}_{\cdot 2}\right)}^{-\hat{\psi}_1} = \overline{Y}_{\cdot 2} - \overline{Y}_{\cdot 3}$$

The analysis of variance (ANOVA) provides a test of the omnibus null hypothesis that $\mu_1 = \mu_2 = \cdots = \mu_p$. This test is equivalent to a simultaneous test of the hypothesis that all possible contrasts among the p means are equal to zero. It is no accident that the between-groups degrees of freedom in a completely randomized ANOVA design is equal to $p - 1$, which is also the number of orthogonal contrasts that can be constructed from p means.

A Priori and a Posteriori Contrasts

In planning an experiment, a researcher usually has in mind a specific set of hypotheses that the experiment is designed to test. Tests that involve these hypotheses are called **a priori** or **planned tests.** This situation can be contrasted with another in which the researcher believes that the treatment affects the dependent variable, and the experiment is designed to accept or reject this notion. If the *F* test of the omnibus null hypothesis is significant, the researcher knows that at least one contrast among the population means is not equal to zero. Interest then turns to determining which contrast or contrasts among the population means is not equal to zero. Tests that are used for **data snooping**—that is, for identifying population contrasts that are not equal to zero following a significant omnibus test—are called **a posteriori, unplanned,** or **post hoc tests.**

Frequently, an experiment involves both a priori and a posteriori tests. After all of the a priori tests have been performed, the researcher may want to test hypotheses suggested by an inspection of the data. The collection of data is often time-consuming and costly. Hence, it is important to extract all information contained in the data. This objective can be accomplished by the judicious use of both a priori and a posteriori tests.

Exploratory versus confirmatory data analysis. A posteriori tests are often used in exploratory data analysis; a priori tests are usually used in confirmatory data analysis. Although both approaches to data analysis have long been used in research, the terms assumed more specialized meanings in the 1970s as a result of John Tukey's work on exploratory techniques and Karl Jöreskog's work on confirmatory techniques. **Exploratory data analysis** is concerned with identifying patterns and features of data and revealing these features. Exploratory techniques are typically used in the preliminary stages of a research program when the researcher does not have sufficient information to make precise predictions or formulate testable models. An important characteristic of the exploratory approach is flexibility in probing the data and responding to patterns that are uncovered in successive stages of the analysis.

Confirmatory data analysis is used after the researcher has accumulated enough information to make predictions or formulate models. The confirmatory approach stresses the evaluation of evidence as compared with the exploratory approach, which stresses the flexible search for evidence. Data that have been collected for a confirmatory analysis should always be subjected to an exploratory analysis. As mentioned earlier, it is important to extract all information contained in the data.

Thus far, I have discussed two issues that are particularly important in selecting a multiple comparison procedure: (1) Are the contrasts orthogonal or nonorthogonal? and (2) Are the contrasts a priori, a posteriori, or a combination of the two? In the following section, I discuss another important factor: the conceptual unit for a Type I error.

Three Kinds of Type I Error Rates

When an experiment involves one contrast, the probability of making a Type I error corresponds to the significance level that is assigned to the contrast. This value, denoted here by α', is usually either .05 or .01. When the experiment involves two or more contrasts, the situation is more complicated. If a researcher tests $C \geq 2$ independent contrasts,[1] each at the α' level of significance, the probability of making one or more Type I errors is

$$\text{(5.1-3)} \qquad \text{Probability of one or more Type I errors} = 1 - (1 - \alpha')^C$$

which is approximately equal to $C \times \alpha'$ for small values of α'. The rationale underlying equation (5.1-3) is as follows. If a contrast is tested at the α' level of significance, the probability of not making a Type I error for that contrast is $1 - \alpha'$. If C independent contrasts are each tested at the α' level of significance, the probability of not making a Type I error for the first, second, . . . , and Cth contrast is, according to the multiplication rule for independent events, the product of the respective probabilities:

$$\overbrace{(1-\alpha')(1-\alpha')\cdots(1-\alpha')}^{C\text{ terms}} = (1-\alpha')^C$$

[1] If the means are normally and independently distributed with mean equal to μ_j and variance equal to $\sigma^2_\varepsilon / n_j$, then the orthogonality of the contrasts is equivalent to the statistical independence of the contrasts.

The expression $(1 - \alpha')^C$ is the probability of *not making* a Type I error for C independent contrasts. The probability of *making* one or more Type I errors is

$$\begin{aligned}\text{Probability of one or more Type I errors} &= 1 - (\text{probability of not making a Type I error for } C \text{ independent contrasts})\\ &= 1 - (1 - \alpha')^C\end{aligned}$$

As the number of independent tests increases, so does the probability of obtaining spuriously significant results. For example, if $\alpha' = .05$ and a researcher tests, say, 3, 5, or 10 independent contrasts, the probability of one or more Type I errors is, respectively,

$$1 - (1 - .05)^3 = .14$$

$$1 - (1 - .05)^5 = .23$$

$$1 - (1 - .05)^{10} = .40$$

For nonindependent tests, the probability of making one or more Type I errors is

$$\text{Probability of one or more Type I errors} \le 1 - (1 - \alpha')^C$$

It is apparent that if enough contrasts are tested, each at the α' level of significance, a researcher will probably reject one or more null hypotheses even though they are all true. An alternative research strategy is to control the Type I error at α for the collection or family of contrasts that are tested. I discuss this strategy next.

A **family of contrasts** consists of those contrasts that are related in terms of their content and intended use. For example, contrasts that involve a control group and two experimental groups are a family. John Tukey (1953) described three kinds of Type I error rates for such contrasts: per-contrast error rate, familywise error rate, and per-family error rate. These error rates are denoted by, respectively, α_{PC}, α_{FW}, and α_{PF}. Suppose that many experiments, each involving a family of contrasts, are performed and we are able to count the number of erroneous conclusions. The three Type I error rates can be defined as follows:

$$\text{Per-contrast error rate } (\alpha_{PC}) = \frac{\text{Number of contrasts falsely declared significant}}{\text{Number of contrasts}}$$

$$\text{Familywise error rate } (\alpha_{FW}) = \frac{\begin{array}{c}\text{Number of families with at least one contrast}\\ \text{falsely declared significant}\end{array}}{\text{Number of families}}$$

$$\text{Per-family error rate } (\alpha_{PF}) = \frac{\text{Number of contrasts falsely declared significant}}{\text{Number of families}}$$

The **per-contrast error rate** is the probability that any one of the contrasts will be incorrectly declared significant. Testing each contrast at the α' level of significance allows the error rate for the family of contrasts to increase as the number of tests increases. An alternative research strategy is to adopt the family of contrasts as the conceptual unit for making a Type I error. If this strategy is adopted, a researcher can choose to control the familywise error rate or the per-family error rate. The **familywise error rate** is the probability of making one or more erroneous statements per family. The **per-family error rate** is the long-run average number of erroneous statements made per family. This error rate is not a probability but rather the expected number of errors per family of contrasts.

An example will help clarify the definitions of the three error rates. Suppose that 1000 replications of an experiment are performed and that for each experiment, 10 contrasts are tested—10,000 tests in all. Also suppose that of the 10,000 tests, 90 tests are incorrectly declared significant, and these 90 incorrect decisions are distributed among 70 of the experiments. The three error rates are as follows:

per comparison error rate

$$\text{Per-contrast error rate} = \frac{\text{Number of contrasts falsely declared significant}}{\text{Number of contrasts}} = \frac{90}{10,000} = .009$$

*

$$\text{Familywise error rate} = \frac{\text{Number of families with at least one contrast falsely declared significant}}{\text{Number of families}} = \frac{70}{1000} = .07$$

$$\text{Per-family error rate} = \frac{\text{Number of contrasts falsely declared significant}}{\text{Number of families}} = \frac{90}{1000} = .09$$

The three Type I error rates become more and more divergent as the number of contrasts in an experiment increases; the three error rates are the same when the experiment involves only one contrast. For $C \geq 2$ independent tests, the relationship among the three error rates is

$$\alpha_{\text{PC}} = \alpha' < \left[\alpha_{\text{FW}} = 1-(1-\alpha')^C\right] < \left[\alpha_{\text{PF}} = \sum_{j=1}^{C} \alpha'\right]$$

If a researcher tests five mutually independent contrasts, each at, say, $\alpha' = .05$, the error rates are

$$\alpha_{\text{PC}} = .05 < \left[\alpha_{\text{FW}} = 1-(1-.05)^5 = .23\right] < \left[\alpha_{\text{PF}} = (5)(.05) = .25\right]$$

For small values of α', the per-family and familywise error rates are numerically almost identical. For example, if a researcher tests five mutually independent contrasts, each at $\alpha' = .01$, the per-family error rate is .05:

$$\alpha_{\text{PF}} = \sum_{j=1}^{5} .01 = .05$$

The familywise error rate is .049:

$$\alpha_{\text{FW}} = 1-(1-.01)^5 = .049$$

* other authors: FWER (Family-wise error rate)
= experimentwise error rate
= P(At least one falsely rejected hypotheses)

If a researcher controls the per-family error rate for $C \geq 2$ tests at α_{PF}, the familywise error rate cannot exceed α_{PF}.

When a completely randomized analysis of variance design is used to test the omnibus null hypothesis

$$H_0\colon \mu_1 = \mu_2 = \cdots = \mu_p$$

the p treatment levels are the conceptual unit for a Type I error. If a test of the omnibus null hypothesis is significant, interest usually shifts to determining which contrasts among the treatment means are significant. It is customary to assign the same error rate to the family of contrasts as was assigned to the omnibus null hypothesis. This principle generalizes to multitreatment ANOVA designs. A factorial design with two treatments involves three tests: treatment A, treatment B, and the $A \times B$ interaction. If, say, a test of treatment A is significant at the $\alpha = .05$ level of significance, it is customary to assign the same $\alpha = .05$ error rate to the family of contrasts associated with treatment A.

For multitreatment ANOVA designs, another conceptual unit for a Type I error can be identified: the experiment. If the familywise Type I error is .05 for treatment A, .05 for treatment B, and .05 for the $A \times B$ interaction, the **experimentwise error rate,** α_{EW}, is equal to

$$\alpha_{\text{EW}} = 1 - (1 - .05)^3 = .143$$

What Is the Correct Conceptual Unit for a Type I Error?

The merits of making the contrast or some larger unit, such as the family or experiment, the conceptual unit for the error rate were extensively debated in the early 1960s. If an experiment involves only one contrast, there is no debate; the error rates for the contrast, family, and experiment are the same. The question only arises when the family or experiment involves two or more contrasts. As you will see, the answer to the question, "What is the correct conceptual unit for a Type I error rate?" depends on the nature of the contrasts of interest.

If orthogonal contrasts have been planned in advance, contemporary practice favors adopting the contrast as the conceptual unit for a Type I error. Earlier you saw that testing a priori, orthogonal contrasts is equivalent to partitioning the data so that each test involves nonredundant pieces of information. Such contrasts are chosen in advance because they address particular research questions of interest. Furthermore, the number of such research questions cannot exceed the number of nonredundant questions, $p - 1$, that can be answered from a set of data. By comparison, nonorthogonal contrasts involve redundant information; the outcome of one test is not independent of those for other tests. Here contemporary practice favors adopting a larger unit such as the family of contrasts as the conceptual unit for a Type I error.

A strong case can be made for these practices. Consider the following two a priori orthogonal contrasts: $\psi_1 = \mu_1 - \mu_2$ and $\psi_2 = (\mu_1 + \mu_2)/2 - \mu_3$. A researcher could choose to conduct a single experiment to test hypotheses about the two contrasts. Alternatively, a researcher could choose to conduct two separate experiments: The first experiment could test the hypothesis $\mu_1 - \mu_2 = 0$ and the second the hypothesis $(\mu_1 + \mu_2)/2 - \mu_3 = 0$. The outcome of the first experiment would provide no information about the probable outcome of the second experiment. This research situation can be contrasted with a second situation in which a researcher is interested in the three pairwise, nonorthogonal contrasts among three means: $\psi_1 = \mu_1 - \mu_2$, $\psi_2 = \mu_1 - \mu_3$, and $\psi_3 = \mu_2 - \mu_3$. Again, the researcher could conduct a single experiment or separate experiments. If the researcher chose to conduct separate experiments to test the three null hypotheses, the reader might anticipate that it would be necessary to conduct *three* separate experiments. Actually, only *two* separate experiments are necessary because the outcome of testing H_0: $\mu_1 - \mu_2 = 0$ and H_0: $\mu_1 - \mu_3 = 0$ could be used to predict the outcome of the third experiment. This follows because, as I showed earlier, $\hat{\psi}_3 = \hat{\psi}_2 - \hat{\psi}_1$:

$$\hat{\psi}_3 = \overbrace{\left(\overline{Y}_{\cdot 1} - \overline{Y}_{\cdot 3}\right)}^{\hat{\psi}_2} - \overbrace{\left(\overline{Y}_{\cdot 1} - \overline{Y}_{\cdot 2}\right)}^{-\hat{\psi}_1} = \overline{Y}_{\cdot 2} - \overline{Y}_{\cdot 3}$$

Contemporary practice treats experiments involving orthogonal contrasts differently from those involving nonorthogonal contrasts. In the first example involving a priori orthogonal contrasts, it is customary to control the per-contrast Type I error rate. In the second example involving nonorthogonal contrasts, contemporary practice favors controlling the familywise or per-family Type I error rate.

The practice of treating orthogonal contrasts differently from nonorthogonal contrasts extends to the analysis of variance. Consider a two-treatment factorial ANOVA design with equal sample sizes in which the researcher has advanced a priori hypotheses about treatments A and B and the $A \times B$ interaction. The two treatments and the interaction represent three orthogonal families of contrasts. The usual practice in analysis of variance is to control the familywise Type I error rather than the experimentwise error.

Complete Versus Partial Null Hypotheses

In choosing a multiple comparison procedure, it is important to consider the nature of the null hypothesis that is to be tested. A null hypothesis can be *complete,* which means that all population means are equal, or *partial,* which means that only a subset of the means is equal. Hayter (1986) recommended that if a researcher wants to control, say, the familywise error rate, a multiple comparison procedure should be chosen that controls the maximum familywise error rate attainable under any complete or partial null hypothesis. Not all multiple comparison procedures meet this requirement. One example is the LSD (least significant difference) multiple comparison procedure proposed by Fisher (1935a), which consists of two steps. In the first step, the omnibus null hypothesis is tested with an analysis of variance F test with $\alpha_{FW} = \alpha'$, where the α' is equal to, say, .05. If the F test is not significant, the omnibus null hypothesis is not rejected, and no more tests are performed. If the omnibus null hypothesis is rejected, Student's t statistic is used to test each pairwise contrast with $\alpha_{PC} = \alpha'$, where the $\alpha' = .05$. Fisher's procedure controls the familywise

Type I error rate when the complete null hypothesis is true. However, if the experiment has more than three treatment levels and the complete null hypothesis is rejected, the familywise error rate exceeds α' (Hayter, 1986). Therefore, Fisher's procedure is not recommended when an experiment has more than three treatment levels because it fails to control the maximum familywise error rate attainable under any complete or partial null hypothesis at a preselected level of significance.

Conceptual Unit for Power

Earlier, I discussed the merits of making the contrast or some larger unit, such as the family or experiment, the conceptual unit for the error rate. A similar issue arises in connection with power. The power of a multiple comparison procedure is the probability of rejecting a false null hypothesis. Other things equal, a researcher wants to use a procedure that both controls the Type I error rate at an acceptable level and provides maximum power. That is easier said than done because there are a number of ways of defining power. One conception of power is **overall power**—the probability of rejecting a false complete null hypothesis. This is the power associated with the *F* test in analysis of variance. Another conception of power, introduced by Einot and Gabriel (1975), is ***P*-subset power.** *P*-subset power focuses on detecting the heterogeneity of means from a subset of a particular size—say, two means or per-pair power, three means or per-triplet power, and so on. Per-pair power, for example, is often expressed as the average probability of detecting true differences among all pairs of means.

In 1978, Ramsey introduced two more conceptions of power: any-pair power and all-pairs power. **Any-pair power** is the probability of detecting at least one true difference among all pairs of means. **All-pairs power** is the probability of detecting all true differences among all pairs of means. There is some debate as to which of the four conceptions of power is more appropriate. Consequently, when researchers investigate the relative power of multiple comparison procedures, it is customary to report data for each kind of power. The different conceptions of power yield different power numbers because any-pair power focuses on the largest mean difference, all-pairs power focuses on the smallest mean difference, and per-pair power is an average that is appropriate for only those two-mean differences that are equal to the average. As would be expected, the any-pair power of multiple comparison procedures is higher than the all-pairs power; the per-pair power falls between that for any-pair power and all-pairs power.

Three Kinds of Test Statistics

Most multiple comparison procedures use one of the following test statistics:

t statistic: $$\frac{\hat{\psi}}{\hat{\sigma}_{\psi}} = \frac{\sum_{j=1}^{p} c_j \overline{Y}_{.j}}{\sqrt{MS_{\text{error}} \sum_{j=1}^{p} \frac{c_j^2}{n_j}}}$$

$$q \text{ statistic:} \qquad \frac{\hat{\psi}}{\hat{\sigma}_{\bar{Y}}} = \frac{\sum_{j=1}^{p} c_j \bar{Y}_{.j}}{\sqrt{\frac{MS_{\text{error}}}{n}}}$$

$$F \text{ statistic:} \qquad \frac{MS_{\text{set of means}}}{MS_{\text{error}}} = \frac{\sum_{j=1}^{s} n_j \bar{Y}_{.j}^2 - \left(\sum_{j=1}^{s} n_j \bar{Y}_{.j}\right)^2 / \sum_{j=1}^{s} n_j}{(s-1)MS_{\text{error}}}$$

where p is the number of means and s is the number of means in a set of the means. The labels *t, q,* and *F* are a convenient way to identify the statistics and their sampling distributions: Student's t distribution, the Studentized range distribution, and the F distribution, respectively. Critical values for these distributions are given in Appendix E. The numerator of the t and q statistics is always a contrast, which is a kind of range; the denominator is either the standard error of a contrast, $\hat{\sigma}_{\psi}$, or the standard error of a mean, $\hat{\sigma}_{\bar{Y}}$. The numerator of an F statistic is computed from all of the means included in a set of means. Both the numerator and the denominator of an F statistic are variances.

For pairwise contrasts with equal sample sizes, the three statistics are related as follows:

$$t = \frac{q}{\sqrt{2}} = \sqrt{F} \qquad \text{or} \qquad \frac{\hat{\psi}}{\hat{\sigma}_{\psi}} = \frac{\hat{\psi}}{\hat{\sigma}_{\bar{Y}}\sqrt{2}} = \sqrt{\frac{MS_{\text{set of means}}}{MS_{\text{error}}}}$$

In general, the F statistic tends to be more powerful than the q statistic, but as you will see, it requires much more computation. Differences among the *t, q,* and *F* statistics are examined in more detail in the following section. Computational examples for the three statistics are given in Sections 5.2 to 5.6.

Single-Step Versus Multiple-Step Procedures

In a 1990 literature survey, I identified more than 30 multiple comparison procedures used by researchers (Kirk, 1990). And the list continues to grow. It is convenient to classify multiple comparison procedures as either single-step or multiple-step procedures. A **single-step procedure** uses one critical value to test hypotheses about contrasts.[2] Any test statistic that exceeds or equals the critical value is declared significant, and the associated null hypothesis is rejected. A variety of single-step procedures are described in Sections 5.2 to 5.7. A **multiple-step procedure** uses two or more critical values to test hypotheses. There are three types of multiple-step procedures: two-step, step-down, and step-up procedures. Fisher's two-step multiple comparison procedure was described earlier in this section. The general features of step-down and step-up procedures are described next.

[2] Single-step procedures also are called simultaneous procedures.

Step-down procedure. Suppose a researcher wants to use a **step-down procedure** to test hypotheses for all pairwise contrasts among five means. The researcher has ordered the means from the smallest to the largest as follows:

$$\bar{Y}_{.1} = 12.9 \qquad \bar{Y}_{.2} = 14.6 \qquad \bar{Y}_{.3} = 16.1 \qquad \bar{Y}_{.4} = 18.8 \qquad \bar{Y}_{.5} = 19.7$$

If a q statistic is used, the step-down procedure begins by testing the contrast involving the smallest and largest means—that is, a contrast in which the means are separated by $r = p$ steps, in this example, five steps. If the null hypothesis $\mu_1 - \mu_5 = 0$ is rejected, then hypotheses $\mu_1 - \mu_4 = 0$ and $\mu_2 - \mu_5 = 0$ are tested. These hypotheses involve means separated by $r = 4$ steps. If these hypotheses are rejected, all hypotheses involving means separated by $r = 3$ steps are tested and finally all means separated by $r = 2$ steps. The critical value that a q statistic must exceed is a function of the number of steps that separate the means. The critical value is largest for contrasts whose means are separated by five steps and smallest for contrasts whose means are separated by two steps. If the null hypothesis for a contrast is not rejected, by implication the null hypotheses for all contrasts encompassed by the nonrejected contrast are not rejected. This testing strategy ensures **coherence.** For example, if a test of the hypothesis $\mu_1 - \mu_3 = 0$ is not rejected, then tests of $\mu_1 - \mu_2 = 0$ and $\mu_2 - \mu_3 = 0$ are not rejected by implication.

When an F statistic is used with a step-down procedure, the first test is the same as an ANOVA F test of the omnibus null hypothesis—that is, a test of $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$. If this hypothesis is rejected, the F statistic is used to test the homogeneity of all subsets of $s = p - 1 = 4$ means—that is, a test of (1) $\mu_1 = \mu_2 = \mu_3 = \mu_4$; (2) $\mu_1 = \mu_2 = \mu_3 = \mu_5$; (3) $\mu_1 = \mu_2 = \mu_4 = \mu_5$; (4) $\mu_1 = \mu_3 = \mu_4 = \mu_5$; and (5) $\mu_2 = \mu_3 = \mu_4 = \mu_5$. Next, the homogeneity of all subsets of three means is tested, excluding those subsets declared homogeneous by implication. Finally, the homogeneity of all subsets of two means is tested, again excluding those subsets declared homogeneous by implication.

It should be evident that more tests are required when an F statistic is used than when a q statistic is used. For example, to reject the hypothesis $\mu_1 - \mu_2 = 0$, the q statistic must have previously rejected $\mu_1 - \mu_5 = 0$, $\mu_1 - \mu_4 = 0$, and $\mu_1 - \mu_3 = 0$—three null hypotheses. To reject the same hypothesis, the F statistic must have previously rejected the homogeneity of (1) $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$, (2) $\mu_1 = \mu_2 = \mu_3 = \mu_4$, (3) $\mu_1 = \mu_2 = \mu_3 = \mu_5$, (4) $\mu_1 = \mu_2 = \mu_4 = \mu_5$, (5) $\mu_1 = \mu_2 = \mu_3$, (6) $\mu_1 = \mu_2 = \mu_4$, and (7) $\mu_1 = \mu_2 = \mu_5$—seven null hypotheses. A computational example using a t statistic is presented in Section 5.2; examples using F and q statistics are presented in Section 5.5.

Step-up procedure. A third type of multiple-step procedure is the **step-up procedure.** Once the p means have been ordered from smallest to largest, hypotheses involving adjacent means are tested. If a null hypothesis for one of these contrasts is rejected, then by implication all null hypotheses that contain the rejected contrast also are rejected. For example, suppose that the null hypothesis $\mu_3 - \mu_4 = 0$ is rejected but $\mu_1 - \mu_2 = 0$, $\mu_2 - \mu_3 = 0$, and $\mu_4 - \mu_5 = 0$ are not rejected. The explicit rejection of $\mu_3 - \mu_4 = 0$ results in the implicit rejection of $\mu_1 - \mu_5 = 0$, $\mu_1 - \mu_4 = 0$, $\mu_2 - \mu_5 = 0$, $\mu_2 - \mu_4 = 0$, and $\mu_3 - \mu_5 = 0$. Because the hypotheses $\mu_1 - \mu_2 = 0$ and $\mu_2 - \mu_3 = 0$, for example, involving adjacent means are not rejected, it is necessary to explicitly test the contrast $\mu_1 - \mu_3 = 0$, which is separated by three steps.

Step-down procedures are widely used in the behavioral sciences, health sciences, and education. Step-up procedures are used less often. Both kinds of procedures tend to be more powerful than single-step procedures. However, step-down and step-up procedures suffer from several shortcomings: (1) In general, they cannot be used to construct confidence intervals; (2) with a few exceptions, they cannot be used to test directional hypotheses; and (3) they tend to require more computation than single-step procedures.

Five Common Hypothesis-Testing Situations

From a review of the literature in the behavioral sciences, health sciences, and education, I have identified five hypothesis-testing situations that occur with some degree of regularity: testing hypotheses about

1. $p - 1$ a priori orthogonal contrasts
2. $p - 1$ a priori nonorthogonal contrasts involving a control group mean
3. C a priori nonorthogonal contrasts
4. All pairwise contrasts among p means
5. All contrasts including nonpairwise contrasts that appear interesting from an inspection of the data

Contrasts in the first category are a priori and orthogonal; those in the other four categories are nonorthogonal. As discussed earlier, for contrasts in the first hypothesis-testing situation, the usual practice is to adopt the individual contrast as the conceptual unit for a Type I error. For the other four hypothesis-testing situations, it is customary to adopt the family of contrasts as the conceptual unit for a Type I error.

Statisticians have developed a variety of test statistics that can be used to control the Type I error rate in these five situations. Table 5.1-1 summarizes the test statistics that I recommend for each situation. The procedures in the upper part of the table assume normality of the population distributions, random sampling or random assignment, and homogeneity of population variances. Tukey's test, the REGW *FQ* test, and the REGW *Q* test also require equal-sized samples. If the assumption of homogeneity of population variances is not tenable or the requirement of equal-sized samples is not met, the multiple comparison procedures in the lower part of Table 5.1-1 can be used. As you will see, the power of the recommended procedures differs markedly. In general, test statistics that were designed for testing a select, limited number of a priori contrasts are more powerful than those designed to test all pairwise comparisons or all possible contrasts. Hence, when possible, it is to a researcher's advantage to specify in advance either orthogonal contrasts or a limited number of contrasts. The problem facing a researcher is to choose the test statistic that provides both the desired kind of Type I error protection and maximum power. The following sections describe the recommended test statistics for each of the five research situations.

Table 5.1-1 ■ Multiple Comparison Procedures That Are Recommended for Five Common Research Situations

Recommended Procedures When Assumptions Are Tenable

	Orthogonal Contrasts	Nonorthogonal Contrasts
A priori contrasts	**1. Testing $p - 1$ contrasts** Student's *t* test (5.2)*	**2. Testing $p - 1$ contrasts with a control group mean** Dunnett's test (5.3)
		3. Testing *C* contrasts** Dunn-Šidák test (5.4) Holm's test (5.4)
A posteriori contrasts		**4. Testing all pairwise contrasts**** Tukey's test (5.5) Fisher-Hayter test (5.5) REGW *F, FQ,* and *Q* tests (5.5)
		5. Testing all contrasts Scheffé's test (5.6)

Recommended Procedures When Assumptions Are Not Tenable

	Orthogonal Contrasts	Nonorthogonal Contrasts
	1. Testing $p - 1$ contrasts: **Heterogeneous variances** Student's t' test with Welch degrees of freedom (5.2)	**2. Testing $p - 1$ contrasts with a control group mean:** **Unequal sample *ns* or heterogeneous variances** Dunnett's test with modifications (5.3)
A priori contrasts		**3. Testing *C* contrasts:** **Heterogeneous variances**** Dunn-Šidák test with Welch degrees of freedom (5.4) Holm's test with Welch degrees of freedom (5.4)

(Continued)

Table 5.1-1 ■ Multiple Comparison Procedures That Are Recommended for Five Common Research Situations (Continued)

A posteriori contrasts		**4. Testing all pairwise contrasts:** **Unequal sample sizes** Tukey-Kramer test (5.5) Fisher-Hayter test (5.5) **Heterogeneous variances**** Dunnett's *T*3 test (5.5) Dunnett's *C* test (5.5) Games-Howell test (5.5)
		5. Testing all contrasts: **Heterogeneous variances** Brown-Forsythe test (5.6)

Note: The recommended procedures control the per-contrast, familywise, or per-family error rate and also have one or a combination of the following virtues: conceptual simplicity, ease of computation, excellent power, availability of confidence intervals, and robustness.

*The numbers in parentheses denote the section in which a procedure is described.

**When more than one procedure is recommended, the procedures are listed in order of increasing power.

5.2 Procedures for Testing $p - 1$ a Priori Orthogonal Contrasts

Student's Multiple *t* Test

Student's t statistic is a single-step procedure that can be used to test null hypotheses of the form

$$H_0\colon \psi_1 = 0$$

$$H_0\colon \psi_2 = 0$$

$$\vdots$$

$$H_0\colon \psi_{p-1} = 0$$

where the $p - 1$ contrasts are a priori and mutually orthogonal. It is not necessary to test the omnibus null hypothesis with an ANOVA F statistic prior to testing the individual contrasts. An omnibus test answers the general question, "Are there any differences among the population means?" If a specific set of orthogonal contrasts has been advanced, a researcher is not interested in this general question. Rather, the researcher is interested in answering a limited number—$p - 1$ or fewer—of specific questions. As I discussed in Section 5.1, current practice favors testing each of the $p - 1$ contrasts at $\alpha_{PC} = \alpha'$, that is, controlling the per-contrast error rate.

The t statistic for testing a null hypothesis is

$$t = \frac{\hat{\psi}_i}{\hat{\sigma}_{\psi_i}} = \frac{\sum_{j=1}^{p} c_j \overline{Y}_{\cdot j}}{\sqrt{MS_{\text{error}} \sum_{j=1}^{p} \frac{c_j^2}{n_j}}} = \frac{c_1\overline{Y}_{\cdot 1} + c_2\overline{Y}_{\cdot 2} + \cdots + c_p\overline{Y}_{\cdot p}}{\sqrt{MS_{\text{error}}\left(\frac{c_1^2}{n_1} + \frac{c_2^2}{n_2} + \cdots + \frac{c_p^2}{n_p}\right)}}$$

where $\hat{\sigma}_{\psi_i}$ is the standard error of the ith contrast and MS_{error} is a pooled estimator of the population error variance. For data that fit a completely randomized ANOVA design, a within-groups mean square (*MSWG*) is used to estimate the population error variance and is given by

$$MSWG = \left[\sum_{j=1}^{p}\sum_{i=1}^{n_j} Y_{ij}^2 - \sum_{j=1}^{p} \frac{\left(\sum_{i=1}^{n_j} Y_{ij}\right)^2}{n_j}\right] \Big/ \sum_{j=1}^{p} (n_j - 1)$$

with $\sum_{j=1}^{p}(n_j - 1)$ degrees of freedom. If the sample sizes are equal, the formula for *MSWG* is

$$MSWG = \left[\sum_{j=1}^{p}\sum_{i=1}^{n} Y_{ij}^2 - \sum_{j=1}^{p} \frac{\left(\sum_{i=1}^{n} Y_{ij}\right)^2}{n}\right] \Big/ p(n-1)$$

with $p(n - 1)$ degrees of freedom. A two-sided null hypothesis is rejected if the absolute value of t exceeds or equals the critical value, $t_{\alpha/2,\,\nu}$, obtained from Student's t distribution in Appendix Table E.3, where α represents the per-contrast error rate and ν is the degrees of freedom associated with the denominator of the t statistic. A one-sided null hypothesis is rejected if the absolute value of t exceeds or equals the critical value, $t_{\alpha,\,\nu}$, and the t statistic is in the predicted tail of the t sampling distribution.

Computational Example Using a t Statistic

The use of Student's t statistic to test hypotheses about a priori orthogonal contrasts is illustrated for an experiment in which 45 subjects have been randomly assigned to five qualitative treatment levels, with 9 subjects in each level. Suppose that the five treatment means are

$$\overline{Y}_{\cdot 1} = 36.7, \overline{Y}_{\cdot 2} = 48.7, \overline{Y}_{\cdot 3} = 43.4, \overline{Y}_{\cdot 4} = 47.2, \overline{Y}_{\cdot 5} = 40.3$$

Assume that the treatment populations are approximately normally distributed and the variances are homogeneous. The design of this experiment corresponds to a completely randomized ANOVA design; hence, *MSWG* is the appropriate estimator of the common population error variance. Assume that the estimate of the population error variance is $MSWG = 29.0322$ with degrees of freedom equal to $p(n-1) = 5(9-1) = 40$. The researcher plans to test the four hypotheses listed in Table 5.2-1. The .05 level of significance is adopted for each test. The reader can verify that the contrasts in Table 5.2-1 are mutually orthogonal.

Table 5.2-1 ■ Coefficients of Orthogonal Contrasts and Associated Statistical Hypotheses

	Coefficients of Contrast					
Contrasts	c_1	c_2	c_3	c_4	c_5	Statistical Hypotheses
ψ_1	0	1	-1	0	0	H_0: $\mu_2 - \mu_3 = 0$ H_1: $\mu_2 - \mu_3 \neq 0$
ψ_2	0	0	0	1	-1	H_0: $\mu_4 - \mu_5 = 0$ H_1: $\mu_4 - \mu_5 \neq 0$
ψ_3	0	$\frac{1}{2}$	$\frac{1}{2}$	$-\frac{1}{2}$	$-\frac{1}{2}$	H_0: $(\mu_2 + \mu_3)/2 - (\mu_4 + \mu_5)/2 = 0$ H_1: $(\mu_2 + \mu_3)/2 - (\mu_4 + \mu_5)/2 \neq 0$
ψ_4	1	$-\frac{1}{4}$	$-\frac{1}{4}$	$-\frac{1}{4}$	$-\frac{1}{4}$	H_0: $\mu_1 - (\mu_2 + \mu_3 + \mu_4 + \mu_5)/4 = 0$ H_1: $\mu_1 - (\mu_2 + \mu_3 + \mu_4 + \mu_5)/4 \neq 0$

The *t* statistics are

$$t = \frac{\hat{\psi}_i}{\hat{\sigma}_{\psi_i}} = \frac{c_1\overline{Y}_{\cdot 1} + c_2\overline{Y}_{\cdot 2} + \cdots + c_p\overline{Y}_{\cdot p}}{\sqrt{MSWG\left(\frac{c_1^2}{n_1} + \frac{c_2^2}{n_2} + \cdots + \frac{c_p^2}{n_p}\right)}}$$

$$t = \frac{\hat{\psi}_1}{\hat{\sigma}_{\psi_1}} = \frac{(1)48.7 + (-1)43.4}{\sqrt{29.0322\left[\frac{(1)^2}{9} + \frac{(-1)^2}{9}\right]}} = \frac{5.300}{2.540} = 2.09$$

$$t = \frac{\hat{\psi}_2}{\hat{\sigma}_{\psi_2}} = \frac{(1)47.2 + (-1)40.3}{\sqrt{29.0322\left[\frac{(1)^2}{9} + \frac{(-1)^2}{9}\right]}} = \frac{6.900}{2.540} = 2.72$$

$$t = \frac{\hat{\psi}_3}{\hat{\sigma}_{\psi_3}} = \frac{(\frac{1}{2})48.7 + (\frac{1}{2})43.4 + (-\frac{1}{2})47.2 + (-\frac{1}{2})40.3}{\sqrt{29.0322\left[\frac{(\frac{1}{2})^2}{9} + \frac{(\frac{1}{2})^2}{9} + \frac{(-\frac{1}{2})^2}{9} + \frac{(-\frac{1}{2})^2}{9}\right]}} = \frac{2.300}{1.796} = 1.28$$

$$t = \frac{\hat{\psi}_4}{\hat{\sigma}_{\psi_4}} = \frac{(1)36.7 + (-\frac{1}{4})48.7 + (-\frac{1}{4})43.4 + (-\frac{1}{4})47.2 + (-\frac{1}{4})40.3}{\sqrt{29.0322\left[\frac{(1)^2}{9} + \frac{(-\frac{1}{4})^2}{9} + \frac{(-\frac{1}{4})^2}{9} + \frac{(-\frac{1}{4})^2}{9} + \frac{(-\frac{1}{4})^2}{9}\right]}} = \frac{-8.200}{2.008} = -4.08$$

The critical value, $t_{.05/2,40}$, required to reject the null hypotheses is 2.021 according to Student's *t* distribution in Appendix Table E.3. Because the absolute value of the *t* statistic for contrasts ψ_1, ψ_2, and ψ_4 exceeds the critical value, the associated null hypotheses can be rejected.

The four *t* tests use the same error mean square (29.0322) in the denominator. As a result, the tests of significance are not statistically independent, even though the contrasts are statistically independent. Research by Norton and Bulgren, as cited by Games (1971), indicates that when the degrees of freedom for MS_{error} are moderately large, say 40, multiple *t* tests can, for all practical purposes, be regarded as independent.

Confidence Intervals for a Priori Orthogonal Contrasts

The *Publication Manual of the American Psychological Association* (American Psychological Association, 2010) states that, in general, reporting confidence intervals is a better strategy than reporting null hypothesis significance tests. I discuss some of the advantages of confidence intervals in Section 2.5. Many of the test statistics in this chapter have confidence interval analogs. Next, I describe a confidence interval analog of Student's multiple *t* statistic.

A 100(1 – α)% confidence interval for an a priori orthogonal contrast is given by

$$\hat{\psi}_i - \hat{\psi}(t) < \psi_i < \hat{\psi}_i + \hat{\psi}(t)$$

where

$$\hat{\psi}_i = c_1\overline{Y}_{\cdot 1} + c_2\overline{Y}_{\cdot 2} + \cdots + c_p\overline{Y}_{\cdot p}$$

$$\hat{\psi}(t) = t_{\alpha/2,\nu}\sqrt{MS_{\text{error}}\left[\frac{c_1^2}{n_1} + \frac{c_2^2}{n_2} + \cdots + \frac{c_p^2}{n_p}\right]}$$

$$\psi_i = c_1\mu_1 + c_2\mu_2 + \cdots + c_p\mu_p$$

Consider the contrast $\psi_1 = \mu_2 - \mu_3$ in Table 5.2-1. The information necessary to construct a 95% confidence interval for this contrast is $\hat{\psi}_1 = 48.7 - 43.4 = 5.3$, $t_{.05/2,\,40} = 2.021$, $MSWG = 29.0322$, and

$$\hat{\psi}(t) = 2.021\sqrt{29.0322\left(\frac{(1)^2}{9} + \frac{(-1)^2}{9}\right)} = 2.021(2.540) = 5.133$$

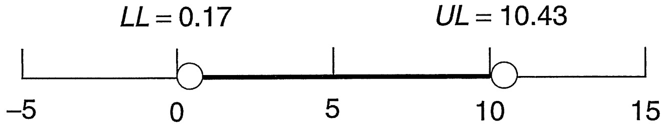
The confidence interval is

$$\hat{\psi}_1 - \hat{\psi}(t) < \psi_1 < \hat{\psi}_1 + \hat{\psi}(t)$$

$$5.3 - 5.133 < \psi_1 < 5.3 + 5.133$$

$$0.17 < \psi_1 < 10.43$$

A researcher can be 95% confident that the open interval [0.17, 10.43] contains the population contrast. This confidence interval corresponds to the darkened portion of the real number line as follows:



95% confidence interval for $\mu_2 - \mu_3$

Confidence intervals permit a researcher to reach the same kind of decision as tests of null hypotheses. Because the interval [0.17, 10.43] does not include zero, a researcher knows that the null hypothesis H_0: $\mu_2 - \mu_3 = 0$ can be rejected. If the confidence interval includes zero, the null hypothesis cannot be rejected. Confidence interval procedures permit a researcher to consider the tenability of all possible null hypotheses, not just the hypothesis that a contrast is equal to zero. For example, the null hypothesis H_0: $\mu_2 - \mu_3 = 12$ would be rejected but not H_0: $\mu_2 - \mu_3 = 8$. Also, the size of the confidence interval provides information about the error variation associated with an estimate and, hence, the strength of the inference. The preference in the *Publication Manual of the American Psychological Association* (American Psychological Association, 2010) for reporting the outcome of confidence intervals rather than hypothesis tests is understandable. Both procedures involve the same assumptions, but confidence intervals provide more information about the data.

Assumptions Associated With Student's *t* Statistic

The assumptions associated with using Student's *t* statistic to test a hypothesis or construct a confidence interval are (1) the observations are drawn from normally distributed populations; (2) the observations are random samples from the populations, or the experimental units are randomly assigned to the treatment levels; and (3) the variances of the populations are equal.

Effects of nonnormality. The effects of sampling from nonnormal populations on the *F* test in analysis of variance are discussed in Section 3.5. Research indicates that if the treatment populations have the same shape—for example, all positively skewed or all leptokurtic—and the sample sizes are fairly large, then the actual probability of making a Type I error is fairly close to the nominal or specified probability. Much less is known about the effects of nonnormality on Student's *t* statistic and on other multiple comparison procedures. Boneau's (1960) research on the *t* statistic and H. J. Keselman and Rogan's (1978) research on a variety of multiple comparison procedures suggest that the results obtained for ANOVA generalize to these procedures. In other words, when sample sizes are large, the *t* statistic and other multiple comparison procedures appear to be robust with respect to nonnormality. This conclusion is consistent with that of Ramseyer and Tcheng (1973). However, the research of Micceri (1989) and Hill and Dixon (1982) on the prevalence of extreme nonnormality in the behavioral sciences, medical sciences, and education is reason for concern.

Effects of heterogeneous variances. The denominator of the *t* statistic uses a pooled estimator of the common population variance. If the population variances are not homogeneous, the use of a pooled estimator may affect the Type I error rate (Games & Howell, 1976). The Type I error is most likely to be affected when the sample sizes are unequal and the smaller samples are obtained from the populations that have the larger variances. Also, Boik (1975) and Kohr and Games (1977) have shown that when the sample sizes or the absolute values of the contrast coefficients are unequal—for example, $\frac{1}{2}$, $\frac{1}{2}$, and 1—the Type I error rate is also likely to be affected by heterogeneous population variances. Under these conditions, one of the robust procedures described next can be used.

Robust Procedures for a Priori Orthogonal Contrasts

If the assumption of the equality of population variances is not tenable, the pooled estimator in the denominator of the *t* statistic can be replaced with individual variance estimators. The resulting statistic, denoted by t', is

$$(5.2\text{-}1) \qquad t' = \frac{\hat{\psi}_i}{\hat{\sigma}_{\psi_i}} = \frac{\sum_{j=1}^{p} c_j\overline{Y}_{\cdot j}}{\sqrt{\sum_{j=1}^{p} \frac{c_j^2\hat{\sigma}_j^2}{n_j}}} = \frac{c_1\overline{Y}_{\cdot 1} + c_2\overline{Y}_{\cdot 2} + \cdots + c_p\overline{Y}_{\cdot p}}{\sqrt{\frac{c_1^2\hat{\sigma}_1^2}{n_1} + \frac{c_2^2\hat{\sigma}_2^2}{n_2} + \cdots + \frac{c_p^2\hat{\sigma}_p^2}{n_p}}}$$

The earliest attempts to determine the sampling distribution of t' were made by Behrens (1929) and enlarged upon by Fisher (1935a). There is no exact solution for this problem. A number of approximate solutions have been proposed: (1) Cochran (1964), (2) Satterthwaite (1946), and (3) Welch (1938, 1947). In general, there is close agreement among these approximate solutions; accordingly, only the approximations of Cochran and Welch are described.

Cochran's procedure uses the t' statistic defined in equation (5.2-1). The two-tailed critical value of t' is given by

$$t'_{\alpha/2,\nu} = \frac{t_{\alpha/2,\nu_j}\left(\dfrac{\hat{\sigma}_j^2}{n_j}\right) + t_{\alpha/2,\nu_{j'}}\left(\dfrac{\hat{\sigma}_{j'}^2}{n_{j'}}\right)}{\dfrac{\hat{\sigma}_j^2}{n_j} + \dfrac{\hat{\sigma}_{j'}^2}{n_{j'}}}$$

where $t_{\alpha/2,\,\nu_j}$ and $t_{\alpha/2,\,\nu_{j'}}$ are the critical values of Student's t distribution at the α level of significance for $\nu_j = n_j - 1$ and $\nu_{j'} = n_{j'} - 1$ degrees of freedom, respectively. The critical value for Cochran's t' is always between the ordinary t values for ν_j and $\nu_{j'}$ degrees of freedom. For a one-tailed test, values of $t_{\alpha,\,\nu_j}$ and $t_{\alpha,\,\nu_{j'}}$ are used. If $n_j = n_{j'}$, then $t' = t$, and the conventional t value with $n_j - 1$ degrees of freedom can be used. The t' test is conservative because the critical value for t' tends to be slightly too large.

Welch's (1938, 1947, 1949) procedure also uses the t' statistic defined in equation (5.2-1). An excellent approximation to the critical value of t' can be obtained from Student's t distribution with degrees of freedom equal to

$$\nu' = \frac{\left(\dfrac{c_1^2\hat{\sigma}_1^2}{n_1} + \dfrac{c_2^2\hat{\sigma}_2^2}{n_2} + \cdots + \dfrac{c_p^2\hat{\sigma}_p^2}{n_p}\right)^2}{\dfrac{c_1^4\hat{\sigma}_1^4}{n_1^2(n_1-1)} + \dfrac{c_2^4\hat{\sigma}_2^4}{n_2^2(n_2-1)} + \cdots + \dfrac{c_p^4\hat{\sigma}_p^4}{n_p^2(n_p-1)}}$$

Wang (1971) reported that when the sample ns are greater than five, Welch's approximate solution controls the Type I error fairly close to α for a wide range of population variances. Similar results were reported by Scheffé (1970). Kohr and Games (1977) reported that Welch's procedure provides reasonable protection against Type I errors when the variances are heterogeneous and the sample sizes or the absolute values of the coefficients of a contrast are unequal.

In summary, when the assumption of the homogeneity of population variances is not tenable, the t' statistic with Welch's modified degrees of freedom is recommended for testing hypotheses about $p - 1$ a priori orthogonal contrasts. Welch's modified degrees of freedom also can be used with other test statistics; I return to this point later.

5.3 Procedures for Testing $p - 1$ Contrasts Involving a Control Group Mean

Dunnett's Multiple Comparison Test

The purpose of many experiments is to compare each of $p - 1$ treatment means with a control group mean. Dunnett (1955) developed a single-step, multiple comparison procedure for this purpose—that is, for testing $p-1$ null hypotheses that have the following form:

$$H_0\colon \psi_1 = \mu_1 - \mu_2 = 0$$

$$H_0\colon \psi_2 = \mu_1 - \mu_3 = 0$$

$$\vdots$$

$$H_0\colon \psi_{p-1} = \mu_1 - \mu_p = 0$$

where μ_1 denotes the control group mean. More specifically, Dunnett's procedure is applicable to any set of $p-1$ a priori nonorthogonal contrasts for which the $p-1$ correlations between the contrasts are equal to 0.5. A correlation of 0.5 occurs, for example, when each of $p-1$ experimental group means is compared with a control group mean and the sample sizes are equal. To illustrate, consider the contrasts in Table 5.3-1, where $\overline{Y}_{.1}$ is the control group mean and the other means are experimental group means. The correlation between the ith and the i'th contrasts is given by

$$\rho_{ii'} = \frac{\sum_{j=1}^{p} c_{ij}c_{i'j} / n}{\sqrt{\left(\sum_{j=1}^{p} c_{ij}^2 / n_j\right)}\sqrt{\left(\sum_{j=1}^{p} c_{i'j}^2 / n_j\right)}}$$

For contrasts $\hat{\psi}_1$ and $\hat{\psi}_2$,

$$\sum_{j=1}^{p} c_{1j}c_{2j} = (1)(1) + (-1)(0) + (0)(-1) + (0)(0) + (0)(0) = 1$$

$$\sum_{j=1}^{p} c_{1j}^2 = (1)^2 + (-1)^2 + (0)^2 + (0)^2 + (0)^2 = 2$$

$$\sum_{j=1}^{p} c_{2j}^2 = (1)^2 + (0)^2 + (-1)^2 + (0)^2 + (0)^2 = 2$$

Suppose that each sample n is equal to 9. The correlation between contrasts $\hat{\psi}_1$ and $\hat{\psi}_2$ is

$$\rho_{12} = \frac{1/9}{\sqrt{(2/9)(2/9)}} = 0.5$$

It can be shown that the correlation between contrast $\hat{\psi}_1$ and the other three contrasts, $\hat{\psi}_3$, $\hat{\psi}_4$, and $\hat{\psi}_5$, also is 0.5.

Dunnett's procedure uses Student's t statistic with equal sample ns. The test statistic is denoted by tDN.

$$tDN = \frac{\hat{\psi}_i}{\hat{\sigma}_{\psi_i}} = \frac{c_j\overline{Y}_{.j} + c_{j'}\overline{Y}_{.j'}}{\sqrt{\dfrac{2MS_{\text{error}}}{n}}}$$

Table 5.3-1 ■ $p - 1$ Contrasts With a Control Group Mean, $\overline{Y}_{.1}$ [Data are from Section 5.2, where $MSWG = 29.0322, p = 5, n = 9$, and $\nu = p(n - 1) = 5(9 - 1) = 40$.]

	Sample Means					
	$\overline{Y}_{.1} = 36.7$	$\overline{Y}_{.2} = 48.7$	$\overline{Y}_{.3} = 43.4$	$\overline{Y}_{.4} = 47.2$	$\overline{Y}_{.5} = 40.3$	
Contrasts	Coefficients of Contrasts					
$\hat{\psi}_1$	1	–1	0	0	0	$\overline{Y}_{.1} - \overline{Y}_{.2} = -12.0$*
$\hat{\psi}_2$	1	0	–1	0	0	$\overline{Y}_{.1} - \overline{Y}_{.3} = -6.7$*
$\hat{\psi}_3$	1	0	0	–1	0	$\overline{Y}_{.1} - \overline{Y}_{.4} = -10.5$*
$\hat{\psi}_4$	1	0	0	0	–1	$\overline{Y}_{.1} - \overline{Y}_{.5} = -3.6$

*$p < .05$; $\hat{\psi}(tDN) = tDN_{.05/2;\,5,\,40}\sqrt{\dfrac{2(29.0322)}{9}} = (2.54)(2.540) = 6.452$ (see text).

A two-sided null hypothesis is rejected if the absolute value of the *tDN* statistic exceeds or equals the critical value $tDN_{\alpha/2;p,\nu}$ obtained from Appendix Table E.7, where α represents the familywise error rate; *p* is the number of treatment means, including the control group mean; and ν is the degrees of freedom associated with the denominator of the *tDN* statistic. A one-sided null hypothesis is rejected if the absolute value of *tDN* exceeds or equals $tDN_{\alpha;\,p,\,\nu}$ and the *tDN* statistic is in the predicted tail of the *tDN* sampling distribution. Dunnett's procedure controls the probability of falsely rejecting one or more null hypotheses—the familywise error rate. It is not necessary to test the omnibus null hypothesis using an ANOVA *F* test prior to using the *tDN* statistic. Indeed, such a test would be pointless.

Instead of computing $p - 1$ test statistics, it is often more convenient to test the $p - 1$ null hypotheses by comparing each contrast with a **critical difference**—a value that the absolute value of a contrast must equal or exceed to be statistically significant. Earlier, I showed that a *tDN* statistic is significant if

$$tDN = \frac{c_j\overline{Y}_{.j} + c_{j'}\overline{Y}_{.j'}}{\sqrt{\dfrac{2MS_{\text{error}}}{n}}} \geq tDN_{\alpha/2;\,p,\,\nu}$$

It follows that the absolute value of any contrast, $|\hat{\psi}_i| = |c_j\overline{Y}_{.j} + c_{j'}\overline{Y}_{.j'}|$, that exceeds or equals

$$\hat{\psi}(tDN) = tDN_{\alpha/2;\,p,\,\nu}\sqrt{\frac{2MS_{\text{error}}}{n}}$$

is statistically significant. The letters $\hat{\psi}(tDN)$ denote the critical difference for Dunnett's procedure. The use of a critical difference to test hypotheses is illustrated for the data in Table 5.3-1, where $\overline{Y}_{\cdot 1}$ is the control group mean. The critical difference that the absolute value of the contrasts in Table 5.3-1 must exceed or equal for a two-tailed test at the $\alpha_{\text{FW}} = .05$ level of significance is

$$\hat{\psi}(tDN) = tDN_{0.5/2;\,5,\,40}\sqrt{\frac{2(29.0322)}{9}} = (2.54)(2.540) = 6.452$$

Because the absolute values of contrasts $\hat{\psi}_1$, $\hat{\psi}_2$, and $\hat{\psi}_3$ exceed the critical difference, the associated null hypotheses can be rejected.

Dunnett's procedure also can be used to establish $p - 1$ simultaneous $100(1 - \alpha)\%$ confidence intervals involving the control group mean. A confidence interval is given by

$$\hat{\psi}_i - \hat{\psi}(tDN) < \psi_i < \hat{\psi}_i + \hat{\psi}_i(tDN)$$

where

$$\hat{\psi}_i = c_j\overline{Y}_{\cdot j} + c_{j'}\overline{Y}_{j'}$$

$$\hat{\psi}(tDN) = tDN_{\alpha/2;\,p,\,\nu}\sqrt{\frac{2MS_{\text{error}}}{n}}$$

$$\psi_i = c_j\mu_j + c_{j'}\mu_{j'}$$

The following assumptions are associated with using Dunnett's *tDN* statistic to test a hypothesis or construct a confidence interval: (1) The observations are drawn from normally distributed populations; (2) the observations are random samples from the populations, or the experimental units are randomly assigned to the treatment levels; (3) the $p - 1$ correlations between contrasts are equal to 0.5; and (4) the variances of the populations are equal. Dunnett (1964) has described modifications of his procedure that can be used when the variance of the control group population is not equal to the variance of the $p - 1$ treatment groups. Hochberg and Tamhane (1987, pp. 140–144) provide tables that can be used when the correlation between two contrasts is not equal to 0.5—a situation that arises when the sample *n*s are not equal.

5.4 Procedures for Testing C a Priori Nonorthogonal Contrasts

In designing an experiment, a researcher usually has a specific set of *C* hypotheses that the experiment is designed to test. Before the research begins, not only is the number of hypotheses known but also which hypotheses are to be tested. Often the associated contrasts are not orthogonal as in comparing a control group mean with $p - 1$ experimental group means or making selected pairwise comparisons among p means. The procedures described in this section can be used to test hypotheses for C a priori nonorthogonal contrasts among p means. A number of multiple comparison procedures have been developed for this purpose. Three are described here: the popular Dunn test, the slightly more powerful Dunn-Šidák test, and the less widely used but more powerful Holm test. The three tests are presented in order of increasing power.

Dunn's Multiple Comparison Test

Fisher described two multiple comparison procedures in his classic experimental design text (Fisher, 1935a, Section 24). One of the procedures bears his name—the Fisher LSD test—and is described in Section 5.1. The originator of the second procedure is unknown. Because Dunn (1961) examined the properties of the second procedure and prepared tables that facilitate its use, the procedure is referred to as **Dunn's multiple comparison procedure.** Some writers use the designation **Bonferroni procedure** because the procedure is based on the Bonferroni or Boole inequality.

Dunn's procedure controls the long-run average number of erroneous statements made per family—the per-family error rate. This is accomplished by dividing α_{PF} into $C \geq 2$ parts: $\alpha_{\text{PF}}/C = \alpha'$. If each of the C contrasts is tested at the α' level of significance, the error rate for the collection of C contrasts is $\alpha_{\text{PF}} = \sum_{i=1}^{C}\alpha'$. For example, if a researcher wants to test $C = 4$ contrasts and wants the per-family error rate to be .05, each contrast can be tested at α' =.05/4 = .0125 level of significance. By testing each contrast at α' = .0125, the per-family error rate is $\alpha_{\text{PF}} = .0125 + .0125 + .0125 + .0125 = .05$.

Dunn developed the procedure using Student's t statistic and sampling distribution. However, the procedure can be used with other test statistics and sampling distributions, which helps to account for its popularity. When Student's t statistic and sampling distribution are used with Dunn's procedure, the statistic is denoted by tD:

$$tD = \frac{\hat{\psi}_i}{\hat{\sigma}_{\psi_i}} = \frac{\sum_{j=1}^{p} c_j\overline{Y}_{\cdot j}}{\sqrt{MS_{\text{error}}\sum_{j=1}^{p}\frac{c_j^2}{n_j}}} = \frac{c_1\overline{Y}_{\cdot 1} + c_2\overline{Y}_{\cdot 2} + \cdots + c_p\overline{Y}_{\cdot p}}{\sqrt{MS_{\text{error}}\left(\frac{c_1^2}{n_1} + \frac{c_2^2}{n_2} + \cdots + \frac{c_p^2}{n_p}\right)}}$$

For pairwise contrasts with equal sample sizes, the statistic simplifies to

$$tD = \frac{\hat{\psi}_i}{\hat{\sigma}_{\psi_i}} = \frac{c_j\overline{Y}_{\cdot j} + c_{j'}\overline{Y}_{\cdot j'}}{\sqrt{\frac{2MS_{\text{error}}}{n}}}$$

A two-sided null hypothesis is rejected if the absolute value of the tD statistic exceeds or equals the critical value $tD_{\alpha/2;\, C,\, \nu}$ obtained from Appendix Table E.14, where α is the per-family error rate, C is the number of contrasts, and ν is the degrees of freedom associated with the denominator of the tD statistic. A one-sided null hypothesis is rejected if the absolute value of tD exceeds or equals $tD_{\alpha;\, C,\, \nu}$ and the tD statistic is in the predicted tail of the tD sampling distribution. Dunn's procedure controls the per-family error rate at α_{PF};

hence, the familywise error rate is less than α_{PF}. It is not necessary to test the omnibus null hypothesis using an ANOVA F test prior to using the tD statistic.

Suppose that a researcher is interested in testing hypotheses for the four nonorthogonal contrasts in Table 5.4-1. The tD test statistics are

$$tD = \frac{\hat{\psi}_i}{\hat{\sigma}_{\psi_i}} = \frac{c_1\overline{Y}_{\cdot 1} + c_2\overline{Y}_{\cdot 2} + \cdots + c_p\overline{Y}_{\cdot p}}{\sqrt{MSWG\left(\dfrac{c_1^2}{n_1} + \dfrac{c_2^2}{n_2} + \cdots + \dfrac{c_p^2}{n_p}\right)}}$$

$$tD = \frac{\hat{\psi}_1}{\hat{\sigma}_{\psi_1}} = \frac{(1)3.67 + (-1)48.7}{\sqrt{29.0322\left[\dfrac{(1)^2}{9} + \dfrac{(-1)^2}{9}\right]}} = \frac{-12.000}{2.540} = -4.72$$

$$tD = \frac{\hat{\psi}_2}{\hat{\sigma}_{\psi_2}} = \frac{(1)3.67 + (-1)43.4}{\sqrt{29.0322\left[\dfrac{(1)^2}{9} + \dfrac{(-1)^2}{9}\right]}} = \frac{-6.700}{2.540} = -2.64$$

$$tD = \frac{\hat{\psi}_3}{\hat{\sigma}_{\psi_3}} = \frac{(1)47.2 + (-1)40.3}{\sqrt{29.0322\left[\dfrac{(1)^2}{9} + \dfrac{(-1)^2}{9}\right]}} = \frac{6.900}{2.540} = 2.72$$

$$tD = \frac{\hat{\psi}_4}{\hat{\sigma}_{\psi_4}} = \frac{\left(\frac{1}{3}\right)36.7 + \left(\frac{1}{3}\right)48.7 + \left(\frac{1}{3}\right)43.4 + \left(-\frac{1}{2}\right)47.2 + \left(-\frac{1}{2}\right)40.3}{\sqrt{29.0322\left[\dfrac{\left(\frac{1}{3}\right)^2}{9} + \dfrac{\left(\frac{1}{3}\right)^2}{9} + \dfrac{\left(\frac{1}{3}\right)^2}{9} + \dfrac{\left(-\frac{1}{2}\right)^2}{9} + \dfrac{\left(-\frac{1}{2}\right)^2}{9}\right]}} = \frac{-0.817}{1.640} = -0.50$$

The critical value, $tD_{.05/2;\,4,\,40}$, required to reject the two-sided null hypotheses is, according to Appendix Table E.14, equal to 2.616. Because the absolute value of the tD statistic for contrasts $\hat{\psi}_1$, $\hat{\psi}_2$, and $\hat{\psi}_3$ exceeds the critical value, the associated null hypotheses can be rejected.

Appendix Table E.14 contains critical values for one- and two-tailed tests. Microsoft's Excel TINV function can be used to obtain critical values for one- and two-tailed tests for any per-family significance level. To obtain a critical value, access the TINV function in Excel,

TINV (probability,deg_freedom)

and replace "probability" with the value of α_{PF}/C for a two-tailed test and with $(2\alpha_{\mathrm{PF}})/C$ for a one-tailed test and "deg_freedom" with the degrees of freedom for MS_{error}. For example, if one-sided null hypotheses had been proposed for the contrasts in Table 5.4-1, the required value of $tD_{.05;\,4,\,40}$ would be given by

TINV[(2)(.05)/4,40] = TINV(.025,40)

and would be 2.329.

Table 5.4-1 ■ Coefficients for C a Priori Nonorthogonal Contrasts [Data are from Section 5.2, where $MSWG = 29.0322, p = 5, n = 9$, and $\nu = p(n-1) = 5(9-1) = 40$.]

	Sample Means					
	$\bar{Y}_{\cdot1} = 36.7$	$\bar{Y}_{\cdot2} = 48.7$	$\bar{Y}_{\cdot3} = 43.4$	$\bar{Y}_{\cdot4} = 47.2$	$\bar{Y}_{\cdot5} = 40.3$	
	Coefficients of Contrasts					
$\hat{\psi}_1$	1	−1	0	0	0	$\bar{Y}_{\cdot1} - \bar{Y}_{\cdot2} = -12.0$
$\hat{\psi}_2$	1	0	−1	0	0	$\bar{Y}_{\cdot1} - \bar{Y}_{\cdot3} = -6.7$
$\hat{\psi}_3$	0	0	0	1	−1	$\bar{Y}_{\cdot4} - \bar{Y}_{\cdot5} = 6.9$
$\hat{\psi}_4$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	$-\frac{1}{2}$	$-\frac{1}{2}$	$\frac{\bar{Y}_{\cdot1} + \bar{Y}_{\cdot2} + \bar{Y}_{\cdot3}}{3} - \frac{\bar{Y}_{\cdot4} + \bar{Y}_{\cdot5}}{2} = -0.8$

Dunn's procedure can be used to establish simultaneous 100(1 − α)% confidence intervals. A confidence interval is given by

$$\hat{\psi}_i - \hat{\psi}(tD) < \psi_i < \hat{\psi}_i + \hat{\psi}(tD)$$

where

$$\hat{\psi}_i = c_1\bar{Y}_{\cdot1} + c_2\bar{Y}_{\cdot2} + \cdots + c_p\bar{Y}_{\cdot p}$$

$$\hat{\psi}_i(tD) = tD_{\alpha/2;\, C,\, \nu}\sqrt{MS_{\text{error}}\left[\frac{c_1^2}{n_1} + \frac{c_2^2}{n_2} + \cdots + \frac{c_p^2}{n_p}\right]}$$

$$\psi_i = c_1\mu_1 + c_2\mu_2 + \cdots + c_p\mu_p$$

The popularity of Dunn's multiple comparison procedure can be attributed to three factors: (1) The procedure provides a simple way to control the per-family and, hence, the familywise Type I error; (2) the concept of dividing α_{PF} among C a priori contrasts is a simple one that can be used with any test statistic; and (3) the procedure is flexible, as the following example illustrates. If a researcher considers the consequences of making a Type I error to be equally serious for all C contrasts, it is reasonable to divide α_{PF} equally among the contrasts. If, however, the consequences of making a Type I error are not equally serious for all C contrasts, α_{PF} can be allocated unequally among the contrasts in a manner that reflects the researcher's a priori concern for Type I and Type II errors. Consider, for example, an experiment involving four contrasts in which the .05

level of significance has been adopted. Instead of testing each contrast at $\alpha' = .05/4 = .0125$, the researcher could allocate α_{PF} as follows: $\alpha'_1 = .02$, $\alpha'_2 = .01$, $\alpha'_3 = .01$, $\alpha'_4 = .01$. The per-family error rate is $\alpha_{\text{PF}} = .02 + .01 + .01 + .01 = .05$, which is the same per-family error rate that would be obtained if α_{PF} were divided equally among the four tests.

As I have shown, Dunn's procedure has a number of desirable properties. The Dunn-Šidák procedure described next shares most of these properties and is slightly more powerful; hence, it is preferred over Dunn's procedure.

Dunn-Šidák Multiple Comparison Test

Dunn's procedure provides an upper bound to the familywise Type I error rate. For small values of α_{PF}, the approximation of the exact familywise Type I error is excellent. However, an even better approximation is provided by a multiplicative inequality proved by Šidák (1967). He showed that the familywise error rate for C nonindependent tests is less than or equal to $\alpha_{\text{FW}} \le 1-(1-\alpha')^C$, which is always less than or equal to $\alpha_{\text{PF}} = \sum_{i=1}^{C}\alpha'$. To control the familywise error rate, each contrast can be tested at the $1-(1-\alpha_{\text{FW}})^{1/C} = \alpha''$ level of significance. For example, suppose a researcher plans to test five nonorthogonal contrasts and wants the familywise Type I error rate to be less than or equal to .05. Use of the additive and multiplicative inequalities results in testing each contrast at, respectively,

Additive inequality	$\alpha' = \alpha_{\text{PF}}/C = .05/5 = .01$
Multiplicative inequality	$\alpha'' = 1-(1-\alpha_{\text{FW}})^{1/C} = 1-(1-.05)^{1/5} = .0102$

Because the Dunn-Šidák procedure is slightly more powerful, it is recommended over Dunn's procedure.

When Student's t statistic and sampling distribution are used with the Dunn-Šidák procedure, the statistic is denoted by *tDS:*

$$tDS = \frac{\hat{\psi}_i}{\hat{\sigma}_{\psi_i}} = \frac{\sum\limits_{j=1}^{p} c_j\overline{Y}_{\cdot j}}{\sqrt{MS_{\text{error}}\sum\limits_{j=1}^{p}\frac{c_j^2}{n_j}}} = \frac{c_1\overline{Y}_{\cdot 1} + c_2\overline{Y}_{\cdot 2} + \cdots + c_p\overline{Y}_{\cdot p}}{\sqrt{MS_{\text{error}}\left(\frac{c_1^2}{n_1} + \frac{c_2^2}{n_2} + \cdots + \frac{c_p^2}{n_p}\right)}}$$

A two-sided null hypothesis is rejected if the absolute value of the *tDS* statistic exceeds or equals the critical value $tDS_{\alpha/2;\,C,\,\nu}$ obtained from Appendix Table E.15, where α denotes the familywise error rate, C is the number of contrasts, and ν is the degrees of freedom associated with the denominator of the *tDS* statistic. A one-sided null hypothesis is rejected if the absolute value of *tDS* exceeds or equals $tDS_{\alpha;\,C,\,\nu}$ and the *tDS* statistic is in the predicted tail of the *tDS* sampling distribution.

For the four contrasts in Table 5.4-1, the values of the Dunn-Šidák test statistic are the same as those for Dunn's procedure:

$$tDS = \frac{\hat{\psi}_1}{\hat{\sigma}_{\psi_1}} = \frac{-12.000}{2.540} = -4.72 \qquad tDS = \frac{\hat{\psi}_2}{\hat{\sigma}_{\psi_2}} = \frac{-6.700}{2.540} = -2.64$$

$$tDS = \frac{\hat{\psi}_3}{\hat{\sigma}_{\psi_3}} = \frac{6.900}{2.540} = 2.72 \qquad tDS = \frac{\hat{\psi}_4}{\hat{\sigma}_{\psi_4}} = \frac{-0.817}{1.640} = -0.50$$

The critical value, $tDS_{.05/2;\,4,\,40}$, required to reject two-sided null hypotheses for these contrasts is 2.608, according to Appendix Table E.15. This critical value is slightly smaller than that for Dunn's procedure, which is $tD_{.05/2;\,4,\,40} = 2.616$. Thus, the Dunn-Šidák procedure is slightly more powerful. Both procedures result in rejecting the null hypotheses for contrasts ψ_1, ψ_2, and ψ_3.

The computation of confidence intervals for the Dunn-Šidák procedure follows that for Dunn's procedure. The term $\hat{\psi}(tD)$ in Dunn's confidence interval is replaced with $\hat{\psi}(tDS)$, where

$$\hat{\psi}(tDS) = tDS_{\alpha/2;\,C,\,\nu}\sqrt{MS_{\text{error}}\left[\frac{c_1^2}{n_1} + \frac{c_2^2}{n_2} + \cdots + \frac{c_p^2}{n_p}\right]}$$

Use $tDS_{\alpha/2;\,C,\,\nu}$ for a two-sided confidence interval and $tDS_{\alpha;\,C,\,\nu}$ for a one-sided confidence interval.

Appendix Table E.15 gives critical values for $\alpha_{FW} = .20, .10, .05$, and .01. Microsoft's Excel TINV function can be use to obtain critical values for any familywise significance level.[3] To obtain critical values, access the TINV function in Excel,

TINV (probability,deg_freedom)

and replace "probability" with the value of $1 - (1 - \alpha_{FW})^{1/C}$ for a two-tailed test and with $2[1 - (1 - \alpha_{FW})^{1/C}]$ for a one-tailed test and "deg_freedom" with the degrees of freedom for MS_{error}. For example, if one-sided null hypotheses had been proposed for the contrasts in Table 5.4-1, the required value of $tDS_{.05;\,4,\,40}$ would be given by

$$\text{TINV}\{2[1 - (1 - .05)^{1/4}],40\} = \text{TINV}(.02548,40)$$

and would be 2.321.

The assumptions associated with using the Dunn and the Dunn-Šidák procedures ar the same as those described earlier for Student's t statistic. Comments about the effects o nonnormality and heterogeneous variances on Student's t statistic also apply to the Dun and Dunn-Šidák statistics. Martin, Toothaker, and Nixon (1989) evaluated 19 multipl

[3] Procedures for obtaining familywise critical values and p values using S-Plus, SAS, and SPSS ar described by Kirk and Hetzer (2006, pp. 149–153).

comparison procedures and concluded that both the Dunn and Dunn-Šidák procedures provide excellent Type I error rate protection when the assumptions of normality and homogeneity of population variances are not tenable. If a researcher is concerned about the heterogeneity of population variances, the t' statistic with Welch's modified degrees of freedom, discussed in Section 5.2, can be used with the Dunn and Dunn-Šidák procedures.

Holm's Sequentially Rejective Bonferroni Test

Holm (1979) proposed a modification that converts Dunn's (Bonferroni) single-step procedure into a more powerful step-down procedure. The modification is quite simple. It consists of ranking the absolute value of C test statistics from the largest to the smallest and then testing the largest test statistic at the $\alpha_1' = \alpha_{\text{PF}}/C$ level of significance, the next largest test statistic at $\alpha_2' = \alpha_{\text{PF}}/(C-1)$, the next largest at $\alpha_3' = \alpha_{\text{PF}}/(C-2), \ldots,$ and the smallest test statistic at $\alpha_C' = \alpha_{\text{PF}}$. The testing procedure terminates when a nonsignificant test statistic is encountered. If the sample sizes are not equal, the test statistics should be ranked on the basis of the p values of the test statistics. Holm showed that the procedure controls the familywise Type I error rate at less than α_{FW}. Holm's procedure is more powerful than Dunn's procedure because it uses a less stringent level of significance for the second through the Cth tests. Recall that Dunn's test is a single-step procedure that uses the same $\alpha' = \alpha_{\text{PF}}/C$ level of significance for all tests.

Holm suggested that a slightly more powerful version of the test could be obtained by using the multiplicative inequality instead of the additive inequality. If the multiplicative inequality is used, the largest test statistic (or smallest p value) is tested at the $\alpha_1'' = 1-(1-\alpha_{\text{FW}})^{1/C}$ level of significance, the next largest at $\alpha_2'' = 1-(1-\alpha_{\text{FW}})^{1/(C-1)}$, the next largest at $\alpha_3'' = 1-(1-\alpha_{\text{FW}})^{1/(C-2)}, \ldots,$ and the smallest test statistic at $\alpha_C'' = \alpha_{\text{FW}}$. Holm's procedure is a general one that can be used with t, q, and F statistics. When Student's t statistic and sampling distribution are used with Holm's procedure, the statistic is denoted by *tH:*

$$tH = \frac{\hat{\psi}_i}{\hat{\sigma}_{\psi_i}} = \frac{\sum\limits_{j=1}^{p} c_j\overline{Y}_{\cdot j}}{\sqrt{MS_{\text{error}} \sum\limits_{j=1}^{p} \dfrac{c_j^2}{n_j}}} = \frac{c_1\overline{Y}_{\cdot 1} + c_2\overline{Y}_{\cdot 2} + \cdots + c_p\overline{Y}_{\cdot p}}{\sqrt{MS_{\text{error}}\left(\dfrac{c_1^2}{n_1} + \dfrac{c_2^2}{n_2} + \cdots + \dfrac{c_p^2}{n_p}\right)}}$$

Critical values of tH for $C_i = C, C-1, C-2, \ldots, 2$ can be obtained from Appendix Table E.15. This table contains critical values for the Dunn-Šidák test, $tDS_{\alpha/2;\, C_i,\, \nu}$. Critical values for $C_i = 1$ can be obtained from Appendix Table E.3, which contains critical values for Student's t test, $t_{\alpha/2,\, \nu}$. A two-sided null hypothesis is rejected if the absolute value of the tH statistic exceeds or equals the critical value $tDS_{\alpha/2;\, C_i,\, \nu}$ or $t_{\alpha/2,\, \nu}$, where α denotes the familywise error rate; C_i is equal to C for the largest test statistic, $C-1$ for the next largest test statistic, . . . , and 1 for the smallest test statistic; and ν is the degrees of freedom associated with the denominator of the tH statistic. A one-sided null hypothesis is rejected if the absolute value of tH exceeds or equals $tDS_{\alpha;\, C_i,\, \nu}$ for $C_i \geq 2$ or $t_{\alpha,\, \nu}$ for $C_i = 1$ and the tH statistic is in the predicted tail of the sampling distribution.

Holm's procedure is illustrated for the four contrasts in Table 5.4-1. The first step in testing the null hypothesis for these contrasts is to rank the absolute values of the test statistics from largest to smallest. The statistics and critical values for the four contrasts are as follows:

Absolute Value of Test Statistics Ranked From Largest to Smallest	Critical Value
$tH = \dfrac{\hat{\psi}_1}{\hat{\sigma}_{\psi_1}} = \dfrac{12.000}{2.540} = 4.72^*$	$tDS_{.05/2;\,4,\,40} = 2.608$
$tH = \dfrac{\hat{\psi}_3}{\hat{\sigma}_{\psi_3}} = \dfrac{6.900}{2.540} = 2.72^*$	$tDS_{.05/2;\,3,\,40} = 2.492$
$tH = \dfrac{\hat{\psi}_2}{\hat{\sigma}_{\psi_2}} = \dfrac{6.700}{2.540} = 2.64^*$	$tDS_{.05/2;\,2,\,40} = 2.323$
$tH = \dfrac{\hat{\psi}_4}{\hat{\sigma}_{\psi_4}} = \dfrac{0.817}{1.640} = 0.50$	$t_{.05/2,\,40} = 2.021$

The asterisks identify significant test statistics. The null hypothesis for contrasts ψ_1, ψ_2, and ψ_3 can be rejected. The critical values required to reject the null hypotheses for the Dunn and Dunn-Šidák procedures are:

Dunn's procedure	$tD_{.05/2;\,4,\,40} = 2.616$
Dunn-Šidák procedure	$tDS_{.05/2;\,4,\,40} = 2.608$

In this example, all three procedures lead to the same decisions. However, Holm's procedure is clearly more powerful than the Dunn and Dunn-Šidák procedures. Because all three procedures control the Type I error rate at or less than α_{FW}, a researcher is advised to use the most powerful procedure, which is Holm's procedure. Holm's procedure shares a disadvantage of most multiple-step procedures: It cannot be used to construct confidence intervals.

The assumptions associated with using Holm's procedure are the same as those described earlier for Student's t statistic. Comments about the effects of nonnormality and heterogeneous variances on Student's t statistic also apply to Holm's procedure. If a researcher is concerned about heterogeneity of population variances, the t' statistic with Welch's modified degrees of freedom, discussed in Section 5.2, can be used with Holm's procedure.

A number of modifications of Holm's procedure have been proposed (Holland & Copenhaver, 1987; Shaffer, 1986). These modifications result in a slight increase in power but at the cost of increased complexity. Tables for implementing Shaffer's modifications of Holm's procedure are provided by Seaman and Serlin (1989). A. Y. Gordon and Salzman (2008) provide a thorough examination of the merits of Holm's procedure relative to other step-down procedures.

5.5 Procedures for Testing All Pairwise Contrasts

Tukey's HSD Test

A variety of procedures have been recommended for testing hypotheses about all pairwise contrasts. Probably the most widely used procedure is the HSD (honestly significant difference) test developed by Tukey (1953). This single-step procedure controls the familywise Type I error rate for the collection of all a posteriori pairwise contrasts. Tukey's HSD test is based on the sampling distribution of the Studentized range, which, like the *t* distribution, was derived by William Sealey Gosset. The letter *q* is used to denote the Studentized range distribution. Tukey's HSD statistic, *qT,* is the ratio of a contrast to the standard error of a mean:

$$qT = \frac{\hat{\psi}_i}{\hat{\sigma}_{\overline{Y}}} = \frac{c_j\overline{Y}_{.j} + c_{j'}\overline{Y}_{.j'}}{\sqrt{\dfrac{MS_{\text{error}}}{n}}}$$

A two-sided null hypothesis is rejected if the absolute value of qT exceeds or equals the critical value $q_{\alpha;\,p,\,\nu}$ obtained from Appendix Table E.6, where α denotes the familywise error rate, p is the number of means in the family, and ν is the degrees of freedom associated with the denominator of the qT statistic. Note that the critical value for Tukey's test, unlike that for the Dunn, Dunn-Šidák, and Holm tests, does not depend on the number of contrasts actually tested but on *p,* the number of means. Tukey's procedure, like all a posteriori procedures, is appropriate for testing only two-sided null hypotheses. The procedure has another limitation: It requires equal sample *n*s. If the sample *n*s are not equal, the Tukey-Kramer procedure, described later, can be used.

Tukey's statistic can be used to test the omnibus null hypothesis, $\mu_1 = \mu_2 = \cdots = \mu_p$, by comparing the largest sample mean with the smallest sample mean. If Tukey's statistic exceeds or equals $q_{\alpha;\,p,\,\nu}$ for this contrast, the omnibus null hypothesis is rejected. Alternatively, the omnibus null hypothesis can be tested using the ANOVA *F* statistic. The omnibus qT test is usually slightly less powerful than the *F* test, although there are some configurations of means for which the qT test is more powerful. For example, a qT test is more likely to reject the omnibus null hypothesis if $p - 2$ of the population means are equal and located halfway between the smallest and largest means. On the other hand, an *F* test is more likely to reject the omnibus null hypothesis if half of the means are equal to the largest mean and the other half are equal to the smallest mean. Both the *F* and qT tests control the familywise error rate at or less than α_{FW}; hence, the more powerful test should be used.

When a researcher wants to test all pairwise contrasts among *p* means and the sample *n*s are equal, it is usually more convenient to compute the one critical difference that each contrast must exceed or equal than it is to compute $p(p - 1)/2$ test statistics. The critical difference, $\hat{\psi}(qT)$, that a pairwise contrast must exceed or equal is given by

$$\hat{\psi}(qT) = q_{\alpha;\,p,\,\nu}\sqrt{\frac{MS_{\text{error}}}{n}}$$

Suppose a researcher wants to test all pairwise contrasts for the data in Table 5.5-1.

Table 5.5-1 ■ Absolute Values of All Pairwise Contrasts Among Means [Data are from Section 5.2, where $MSWG = 29.0322$, $p = 5$, $n = 9$, and $\nu = p(n - 1) = 5(9 - 1) = 40$. The means in the table are ordered from the smallest to the largest so that the absolute value of the largest contrast, 12.0, appears in the upper right corner of the table.]

	$\bar{Y}_{\cdot 1} = 36.7$	$\bar{Y}_{\cdot 5} = 40.3$	$\bar{Y}_{\cdot 3} = 43.4$	$\bar{Y}_{\cdot 4} = 47.2$	$\bar{Y}_{\cdot 2} = 48.7$
$\bar{Y}_{\cdot 1} = 36.7$	—	3.6	6.7	10.5*	12.0*
$\bar{Y}_{\cdot 5} = 40.3$		—	3.1	6.9	8.4*
$\bar{Y}_{\cdot 3} = 43.4$			—	3.8	5.3
$\bar{Y}_{\cdot 4} = 47.2$				—	1.5
$\bar{Y}_{\cdot 2} = 48.7$					—

$^*p < .05$; $\hat{\psi}(qT) = q_{05;\,5,\,40}\sqrt{\frac{29.0322}{9}} = (4.04)(1.796) = 7.26.$

The critical difference is

$$\hat{\psi}(qT) = q_{.05;\,5,\,40}\sqrt{\frac{29.0322}{9}} = (4.04)(1.796) = 7.26$$

It is often convenient to construct a table like Table 5.5-1 that gives the absolute value of all pairwise contrasts. Any contrast that exceeds or equals the critical difference is declared significant. It is apparent from Table 5.5-1 that three contrasts exceed the critical difference; hence, the null hypotheses for these contrasts can be rejected.

Tukey's procedure can be used to establish $100(1 - \alpha)\%$ simultaneous confidence intervals for all pairwise population contrasts. The confidence interval is given by

$$\hat{\psi}_i - \hat{\psi}(qT) < \psi_i < \hat{\psi}_i + \hat{\psi}(qT)$$

Earlier I noted that Tukey's procedure requires equal-size samples. In addition, the procedure assumes that (1) the observations are drawn from normally distributed populations; (2) the observations are random samples from the populations, or the experimental units are randomly assigned to the treatment levels; and (3) the variances of the populations are equal. The next two sections describe procedures for testing all pairwise contrasts that do not require equal sample sizes or the assumption that the population variances are equal.

Procedures for Unequal Sample Sizes

Researchers in the social sciences and education frequently want to test all pairwise contrasts among means. Many of the most popular multiple comparison procedures used for this purpose, such as Tukey's HSD test, require equal sample sizes. Unfortunately, researchers live in an imperfect world in which unequal sample sizes are the rule rather than the exception. This dilemma has sparked a search for alternative procedures that can be used when sample sizes are unequal. Most of the research has focused on finding alternatives to Tukey's HSD test. Suggested alternatives include Gabriel's (1978) test, Genizi and Hochberg's (1978) test, Hochberg's (1974) GT2 test, Hunter's (1976) *H* test, Spjøtvoll and Stoline's (1973) *T′* test, and the Tukey-Kramer test (Kramer, 1956; Tukey, 1953). The results of numerous studies of the various alternatives are clear cut: The preferred procedure is the Tukey-Kramer test. This procedure controls the Type I error at less than α_{FW} and has the highest power of the procedures investigated (Dunnett, 1980a; Hayter, 1984; Stoline, 1981).

Tukey-Kramer test. The Tukey-Kramer test was independently proposed by Tukey (1953) and Kramer (1956) for the case in which the sample *n*s are unequal and the basic assumptions of normality, homogeneity of variances, and so on are tenable. The test statistic, denoted by *qTK,* is

$$qTK = \frac{\hat{\psi}_i}{\hat{\sigma}_{\overline{Y}}} = \frac{c_j\overline{Y}_{.j} + c_{j'}\overline{Y}_{.j'}}{\sqrt{\left[MS_{\text{error}}\left(\dfrac{1}{n_j} + \dfrac{1}{n_{j'}}\right)\right]/2}}$$

A two-sided null hypothesis is rejected if the absolute value of *qTK* exceeds or equals the critical value $q_{\alpha;\,p,\,\nu}$ obtained from the Studentized range distribution in Appendix Table E.6, where α denotes the familywise error rate, *p* is the number of means in the family, and ν is the degrees of freedom associated with the denominator of the *qTK* statistic.

Procedures for Heterogeneous Variances

A variety of procedures have been proposed for testing hypotheses about all pairwise contrasts among *p* means when the population variances are heterogeneous. The leading contenders are Dunnett's (1980b) *T*3 and *C* tests and the Games-Howell *GH* test (Games & Howell, 1976). All three tests can be used when the sample sizes are unequal. However, if the population variances are homogeneous, the Tukey-Kramer procedure is recommended because of its superior power.

Dunnett's *T*3 test. The test statistic for Dunnett's *T*3 procedure, denoted by $mT3$, is

$$mT3 = \frac{c_j\overline{Y}_{\cdot j} + c_{j'}\overline{Y}_{\cdot j'}}{\sqrt{\dfrac{\hat{\sigma}_j^2}{n_j} + \dfrac{\hat{\sigma}_{j'}^2}{n_{j'}}}}$$

A two-sided null hypothesis is rejected if the absolute value of $mT3$ exceeds or equals the critical value $m_{\alpha;\,C,\,\nu'}$ obtained from the Studentized maximum modulus distribution in Appendix Table E.16, where α denotes the familywise error rate, $C = p(p - 1)/2$, and ν' denotes the use of Welch's modified degrees of freedom, discussed in Section 5.2:

$$\nu' = \frac{\left(\dfrac{\hat{\sigma}_j^2}{n_j} + \dfrac{\hat{\sigma}_{j'}^2}{n_{j'}}\right)^2}{\dfrac{\hat{\sigma}_j^4}{n_j^2(n_j - 1)} + \dfrac{\hat{\sigma}_{j'}^4}{n_{j'}^2(n_{j'} - 1)}}$$

Dunnett's *C* test. The test statistic for Dunnett's *C* procedure, denoted by qC, is

$$qC = \frac{c_j\overline{Y}_{\cdot j} + c_{j'}\overline{Y}_{\cdot j'}}{\sqrt{\left(\dfrac{\hat{\sigma}_j^2}{n_j} + \dfrac{\hat{\sigma}_{j'}^2}{n_{j'}}\right)/2}}$$

A two-sided null hypothesis is rejected if the absolute value of qC exceeds or equals the critical value

$$qC_{\alpha;p,\,\nu} = \frac{q_{\alpha;p,\,\nu_j}\left(\dfrac{\hat{\sigma}_j^2}{n_j}\right) + q_{\alpha;p,\,\nu_{j'}}\left(\dfrac{\hat{\sigma}_{j'}^2}{n_{j'}}\right)}{\dfrac{\hat{\sigma}_j^2}{n_j} + \dfrac{\hat{\sigma}_{j'}^2}{n_{j'}}}$$

where $q_{\alpha;p,\,\nu_j}$ is obtained from the Studentized range distribution, α denotes the familywise error rate, p is the number of means in the family, and ν_j is equal to $n_j - 1$. This critical value is based on Cochran's (1964) approximate solution to the Behrens-Fisher problem discussed in Section 5.2.

Games-Howell test. The test statistic for the Games-Howell procedure, denoted by qGH, is

$$qGH = \frac{\hat{\psi}}{\hat{\sigma}_{\overline{Y}}} = \frac{c_j\overline{Y}_{\cdot j} + c_{j'}\overline{Y}_{\cdot j'}}{\sqrt{MS_{\text{error}}\left(\dfrac{1}{n_j} + \dfrac{1}{n_{j'}}\right)/2}}$$

A two-sided null hypothesis is rejected if the absolute value of qGH exceeds or equals the critical value $q_{\alpha;\, p,\, \nu'}$ obtained from the Studentized range distribution in Appendix Table E.6, where α denotes the familywise error rate, p is the number of means in the family, and ν' denotes the use of Welch's modified degrees of freedom. The formula for ν' is the same as that given earlier for Dunnett's *T*3 procedure.

The relative merits of these multiple comparison procedures and others that have been recommended for the case of heterogeneous variances and unequal sample sizes have been investigated by a number of researchers (Dunnett, 1980b; Games, Keselman, & Rogan, 1981; H. J. Keselman & Rogan, 1978; Tamhane, 1979). The results of these investigations can be summarized as follows:

1. The Games-Howell procedure is always more powerful than Dunnett's *C* procedure. However, the Games-Howell procedure becomes liberal with respect to the familywise Type I error rate as the variances become more similar. The procedure is the clear choice if the population variances are known to be unequal.
2. The *C* procedure is more powerful than the *T*3 procedure when the number of error degrees of freedom is large and less powerful when the number is small. Both procedures control the familywise Type I error rate. Either the *T*3 or the *C* procedure is a better choice than the Games-Howell procedure if control of the familywise Type I error rate is especially important and the population variances are believed to be similar.

Fortunately, the use of these procedures, particularly the Games-Howell procedure, does not lead to a substantial loss of power relative to procedures that assume equal variances.

Fisher-Hayter Test

Hayter (1986) proposed a modification of Fisher's LSD test that can be used to test hypotheses about all pairwise contrasts. This two-step procedure, which assumes equal variances, controls the familywise Type I error at α_{FW} when the sample *n*s are equal or when the sample *n*s are unequal and the number of means is $p = 3$. When the sample *n*s are unequal and $p > 3$, the Type I error rate cannot exceed α_{FW}. The procedure, which is called the **Fisher-Hayter test,** has two steps. In the first step, the omnibus null hypothesis is tested at the $\alpha'' = \alpha_{\text{FW}}$ significance level using either an F or a q statistic. The critical values for F and q are denoted by, respectively, $F_{\alpha;\, \nu_1,\, \nu_2}$ and $q_{\alpha;\, p,\, \nu}$, where $q_{\alpha;\, p,\, \nu}$ is obtained from Appendix Table E.6 and ν denotes the degrees of freedom for MS_{error}. If this test is not significant, the omnibus null hypothesis is not rejected and no more tests are performed. If the omnibus null hypothesis is rejected, each of the pairwise contrasts is tested at the $\alpha'' = \alpha_{\text{FW}}$ significance level using

$$qFH = \frac{\hat{\psi}}{\hat{\sigma}_{\overline{Y}}} = \frac{c_j\overline{Y}_{\cdot j} + c_{j'}\overline{Y}_{\cdot j'}}{\sqrt{\left[MS_{\text{error}}\left(\dfrac{1}{n_j} + \dfrac{1}{n_{j'}}\right)\right] / 2}}$$

A two-sided null hypothesis is rejected if the absolute value of qFH exceeds or equals the critical value $q_{\alpha;\,p-1,\,\nu}$ obtained from the Studentized range distribution in Appendix Table E.6, and α denotes the familywise Type I error. Note that the table is entered for $p - 1$ means instead of p means. The Fisher-Hayter procedure shares a disadvantage of most multiple-step procedures: It cannot be used to construct confidence intervals. The assumptions associated with the procedure are the same as those described earlier for Student's t statistic (see Section 5.2).

When a researcher wants to test all pairwise contrasts among p means and the sample ns are equal, it is usually more convenient to compute the one critical difference that each contrast must exceed or equal than it is to compute $p(p-1)/2$ test statistics. If the omnibus null hypothesis is rejected, the critical difference, $\hat{\psi}(qFH)$, that a pairwise contrast must exceed or equal is given by

$$\hat{\psi}(qFH) = q_{\alpha;\,p-1,\,\nu}\sqrt{\frac{MS_{\text{error}}}{n}}$$

For the data in Table 5.5-1, the critical difference is

$$\hat{\psi}(qFH) = q_{.05;\,4,\,40}\sqrt{\frac{29.0322}{9}} = (3.79)(1.796) = 6.81$$

With this critical difference, four contrasts in Table 5.5-1 can be rejected—one more than was rejected using Tukey's critical difference. This result is not surprising. Seaman, Levin, and Serlin (1991) compared 23 multiple comparison procedures in terms of familywise Type I error protection and power. They concluded that the Fisher-Hayter test was just slightly less powerful than the most powerful procedures—the REGW and Peritz tests—and represented an excellent trade-off between power and ease of application.

It is instructive to compare the critical difference for the Fisher-Hayter procedure, $\hat{\psi}(qFH) = 6.807$, with those for the Tukey, Dunn, and Dunn-Šidák procedures. When the procedures are used to make all 10 pairwise comparisons among the five means, the critical differences that each contrast must exceed to be significant are

Fisher-Hayter critical difference $\quad \hat{\psi}(qFH) = qFH_{.05;\,4,\,40}\sqrt{\frac{29.0322}{9}} = (3.79)(1.796) = 6.81$

Tukey critical difference $\quad \hat{\psi}(qT) = q_{.05;\,5,\,40}\sqrt{\frac{29.0322}{9}} = (4.04)(1.796) = 7.26$

Dunn critical difference $\quad \hat{\psi}(tD) = tD_{.05/2;\,10,\,40}\sqrt{\frac{2(29.0322)}{9}} = (2.971)(2.540) = 7.55$

Dunn-Šidák critical difference $\quad \hat{\psi}(tDS) = tDS_{.05/2;\,10,\,40}\sqrt{\frac{2(29.0322)}{9}} = (2.963)(2.540) = 7.53$

It is apparent that when all pairwise contrasts are tested, the Fisher-Hayter procedure is more powerful than the other procedures. However, the Dunn and Dunn-Šidák procedures become more powerful relative to the Fisher-Hayter procedure as the number of comparisons among the p means is reduced. For example, if a researcher had planned to make only 4 instead of all 10 pairwise comparisons, the critical difference for the Dunn-Šidák procedure would have been

$$\hat{\psi}(tDS) = tDS_{.05/2;\,4,\,40}\sqrt{\frac{2(29.0322)}{9}} = (2.608)(2.540) = 6.62$$

which is less than the 6.81 required for the Fisher-Hayter procedure.

Holm's procedure also is more powerful than the Fisher-Hayter procedure if only four pairwise contrasts are tested. The critical differences for four contrasts are as follows:

$$\begin{aligned}
\hat{\psi}(tH) &= tDS_{.05/2;\,4,\,40}\sqrt{\frac{2(29.0322)}{9}} = (2.608)(2.540) = 6.62\\
\hat{\psi}(tH) &= tDS_{.05/2;\,3,\,40}\sqrt{\frac{2(29.0322)}{9}} = (2.492)(2.540) = 6.33\\
\hat{\psi}(tH) &= tDS_{.05/2;\,2,\,40}\sqrt{\frac{2(29.0322)}{9}} = (2.323)(2.540) = 5.90\\
\hat{\psi}(tH) &= t_{.05/2,\,40}\sqrt{\frac{2(29.0322)}{9}} = (2.021)(2.540) = 5.13
\end{aligned}$$

The point of this discussion is that for a priori contrasts, a researcher should carefully consider which multiple comparison procedure provides the desired Type I error protection and maximizes power.

REGW *F*, *FQ*, and *Q* Tests

T. A. Ryan (1959, 1960) proposed a step-down multiple comparison procedure for testing hypotheses for all pairwise contrasts. The procedure can be used with either F or q statistics. His procedure, which is more powerful than the Tukey and Fisher-Hayter procedures, uses adjusted significance levels denoted by α'_r. To use Ryan's procedure, the means are ordered from the smallest to the largest mean. A contrast involving the smallest and largest means is said to be separated by $r = p$ steps (the number of means). This contrast is tested at the $\alpha'_r = \alpha_{\text{PF}}(r/p)$ level of significance, where $r = p$. If and only if the contrast is significant, the two contrasts involving means separated by $r = p - 1$ steps are tested at the $\alpha'_r = \alpha_{\text{PF}}(r/p)$ level of significance, and so on. Consider an example with $p = 5$ means; let $\alpha_{\text{FW}} = .05$. Means separated by $r = 5$ steps are tested at the $\alpha'_5 = .05(5/5) = .05$ level of significance, means separated by $r = 4$ steps are tested at the $\alpha'_4 = .05(4/5) = .04$ level of significance, . . . , and means separated by $r = 2$ steps are tested at the $\alpha'_2 = .05(2/5) = .02$ level of significance. If the null hypothesis for a contrast is not rejected, by implication the null hypotheses for all contrasts encompassed by the nonrejected contrast are not rejected.

The advantage of Ryan's procedure is that it controls the familywise Type I error at less than α_{PF} and has greater power than procedures that use a uniform level of significance, such as Tukey's procedure.

Einot and Gabriel (1975) used Ryan's idea of adjusted significance levels but replaced the Bonferroni additive inequality with the multiplicative inequality; means separated by r steps are tested at the $\alpha_r'' = 1 - (1 - \alpha_{\text{FW}})^{r/p}$ level of significance. This modification results in slightly more powerful tests. For example, means separated by

Table 5.5-2 ■ Computational Procedures for the REGW *FQ* Test [Data are from Table 5.5-1, where $MSWG = 29.0322$, $p = 5$, $n = 9$, and $\nu = p(n - 1) = 5(9 - 1) = 40$. To simplify the presentation, the means have been relabeled so that $\overline{Y}_{.1}$ denotes the smallest mean and $\overline{Y}_{.5}$ denotes the largest mean: $\overline{Y}_{.1} = 36.7$, $\overline{Y}_{.2} = 40.3$, $\overline{Y}_{.3} = 43.4$, $\overline{Y}_{.4} = 47.2$, $\overline{Y}_{.5} = 48.7$.]

Number of Means	Hypothesis	*F* Statistic	REGW Critical Value and Decision*
5	$\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$	7.51	$F_{.05;\,4,\,40} = 2.61$ S
Number of Steps Between Means	**Hypothesis**	**| *q* Statistic |**	**REGW Critical Value and Decision***
5	$\mu_1 - \mu_5 = 0$	6.68	$q_{.05;\,4,\,40} = 3.79$ S
4	$\mu_1 - \mu_4 = 0$	5.85	$q_{.05;\,4,\,40} = 3.79$ S
4	$\mu_2 - \mu_5 = 0$	4.68	S
3	$\mu_1 - \mu_3 = 0$	3.73	$q_{.0303;\,3,\,40} = 3.73$ S
3	$\mu_2 - \mu_4 = 0$	3.84	S
3	$\mu_3 - \mu_5 = 0$	2.95	NS
2	$\mu_1 - \mu_2 = 0$	2.00	$q_{.0203;\,2,\,40} = 3.40$ NS
2	$\mu_2 - \mu_3 = 0$	1.73	NS
2	$\mu_3 - \mu_4 = 0$	2.12	NSI
2	$\mu_4 - \mu_5 = 0$	.84	NSI

*S = significant; NS = not significant; NSI = not significant by implication.

four steps are tested at the $\alpha''_4 = 1 - (1 - .05)^{4/5} = .0402$ level of significance instead of the $\alpha'_4 = .05(4/5) = .04$ level. Welsch (1977) further improved the procedure by showing that means separated by $r = p - 1$ steps also could be tested at the same α''_r level as means separated by $r = p$ steps and still control the familywise Type I error at less than α_{FW}. Ryan's idea of adjusted significance levels has undergone numerous modifications and has appeared under a number of different names.[4] To give the major contributors—Ryan, Einot, Gabriel, and Welsch—their just due, the procedure is referred to as the REGW procedure. The designations REGW *F* and REGW *Q* are used to distinguish between the *F* and *q* versions of the test.

Shaffer (1979) proposed yet another improvement on Ryan's idea of adjusted significance levels that can be used with any step-down *q* procedure. The improvement consists of first performing an omnibus ANOVA *F* test on the *p* means. If the *F* test is not significant, the testing sequence terminates. If the *F* test is significant, means separated by $r = p$ steps are tested using the *q* critical value appropriate for means separated by $r = p - 1$ steps. Subsequent tests are performed with the usual *q* critical values. This procedure is referred to as the REGW *FQ* procedure.

The REGW *FQ* procedure is illustrated in Table 5.5-2. The procedure begins with an *F* test of the omnibus null hypothesis. The *F* test statistic is

$$F = \frac{MS_{\text{set of means}}}{MS_{\text{error}}} = \frac{\sum_{j=1}^{s} n_j\overline{Y}_{\cdot j}^2 - \left(\sum_{j=1}^{s} n_j\overline{Y}_{\cdot j}\right)^2 / \sum_{j=1}^{s} n_j}{(s-1)MSWG} = \frac{872.3880}{(5-1)29.0322} = 7.51$$

where $s = 5$. The critical value is $F_{\alpha;\, \nu_1,\, \nu_2} = F_{.05;\, 4,\, 40} = 2.61$; hence, the omnibus null hypothesis is rejected. Following the rejection of the omnibus null hypothesis, all pairwise contrasts are tested using the *qREGW* test statistic:

$$qREGW = \frac{\hat{\psi}_i}{\hat{\sigma}_{\psi_i}} = \frac{c_j\overline{Y}_{\cdot j} + c_{j'}\overline{Y}_{\cdot j'}}{\sqrt{\dfrac{MS_{\text{error}}}{n}}}$$

A two-sided null hypothesis is rejected if (1) the absolute value of *qREGW* exceeds or equals the critical value $q_{\alpha''_r;\, r,\, \nu}$ obtained from the Studentized range distribution in

[4] The following names have been used: (1) modified Ryan's test (Jaccard, Becker, & Wood, 1984), (2) REGWF and REGWQ tests (SAS Institute, Inc., 1985), (3) revised Ryan's procedure (Ramsey, 1978), (4) Ryan's procedure (Einot & Gabriel, 1975; Ramsey, 1981), and (5) Tukey-Welsch procedure (Hochberg & Tamhane, 1987, p. 69; Lehman & Shaffer, 1979).

Table 5.5-3 ■ Computational Procedures for the REGW F Test [Data are from Table 5.5-1, where $MSWG = 29.0322$, $p = 5$, $n = 9$, and $\nu = p(n - 1) = 5(9 - 1) = 40$. The means have been relabeled so that $\bar{Y}_{\cdot 1}$ denotes the smallest mean and $\bar{Y}_{\cdot 5}$ denotes the largest mean: $\bar{Y}_{\cdot 1} = 36.7$, $\bar{Y}_{\cdot 2} = 40.3$, $\bar{Y}_{\cdot 3} = 43.4$, $\bar{Y}_{\cdot 4} = 47.2$, $\bar{Y}_{\cdot 5} = 48.7$.]

Number of Means	Hypothesis	F Statistic	REGW Critical Value and Decision*
			$F_{.05;\,4,\,40} = 2.61$
5	$\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$	7.51	S
			$F_{.05;\,3,\,40} = 2.84$
4	$\mu_1 = \mu_2 = \mu_3 = \mu_4$	6.19	S
	$\mu_1 = \mu_2 = \mu_3 = \mu_5$	8.01	S
	$\mu_1 = \mu_2 = \mu_4 = \mu_5$	10.01	S
	$\mu_1 = \mu_3 = \mu_4 = \mu_5$	8.88	S
	$\mu_2 = \mu_3 = \mu_4 = \mu_5$	4.46	S
			$F_{.0303;\,2,\,40} = 3.82$
3	$\mu_1 = \mu_2 = \mu_3$	3.49	NS
	$\mu_1 = \mu_2 = \mu_4$	8.83	S
	$\mu_1 = \mu_2 = \mu_5$	11.76	S
	$\mu_1 = \mu_3 = \mu_4$	8.76	S
	$\mu_1 = \mu_3 = \mu_5$	11.21	S
	$\mu_1 = \mu_4 = \mu_5$	13.25	S
	$\mu_2 = \mu_3 = \mu_4$	3.70	NS
	$\mu_2 = \mu_3 = \mu_5$	5.59	S
	$\mu_2 = \mu_4 = \mu_5$	6.22	S
	$\mu_3 = \mu_4 = \mu_5$	2.31	NS
			$F_{.0203;\,1,\,40} = 5.84$
2	$\mu_1 = \mu_2$	2.01	NSI
	$\mu_1 = \mu_3$	6.96	NSI
	$\mu_1 = \mu_4$	17.09	S
	$\mu_1 = \mu_5$	22.32	S
	$\mu_2 = \mu_3$	1.49	NSI
	$\mu_2 = \mu_4$	7.38	NSI
	$\mu_2 = \mu_5$	10.94	S
	$\mu_3 = \mu_4$	2.24	NSI
	$\mu_3 = \mu_5$	4.35	NSI
	$\mu_4 = \mu_5$	0.35	NSI

*S = significant; NS = not significant; NSI = not significant by implication.

Appendix Table E.6 and (2) the means in the hypothesis are not encompassed by a nonrejected hypothesis at an earlier stage in the testing procedure. The critical value, $q_{\alpha_r''; r, \nu}$, for means separated by $r = 5$ steps is $q_{.05,\,4,\,40} = 3.79$, where $\alpha_5'' = 1 - (1 - .05)^{5/5} = .05$ (Einot and Gabriel's contribution), $r = 5 - 1 = 4$ (Shaffer's contribution), and $\nu = 40$. The remaining critical values are as follows:

Means separated by $r = 4$ steps: The critical value is $q_{.05,\,4,\,40} = 3.79$, where $\alpha_4'' = 1 - (1 - .05)^{5/5} = .05$ (Welsch's contribution), $r = 5 - 1 = 4$, and $\nu = 40$.

Means separated by $r = 3$ steps: The critical value is $q_{.0303,\,3,\,40} = 3.73$, where $\alpha_3'' = 1 - (1 - .05)^{3/5} = .0303$, $r = 5 - 2 = 3$, and $\nu = 40$.

Means separated by $r = 2$ steps: The critical value is $q_{.0203,\,2,\,40} = 3.40$, where $\alpha_2'' = 1 - (1 - .05)^{2/5} = .0203$, $r = 5 - 3 = 2$, and $\nu = 40$.

In Table 5.5-2, the following pairwise null hypotheses are rejected: $\mu_1 - \mu_5 = 0$, $\mu_1 - \mu_4 = 0$, $\mu_2 - \mu_5 = 0$, $\mu_1 - \mu_3 = 0$, and $\mu_2 - \mu_4 = 0$. The hypotheses $\mu_3 - \mu_4 = 0$ and $\mu_4 - \mu_5 = 0$ are not significant by implication because they are encompassed by the hypothesis $\mu_3 - \mu_5 = 0$, which was not rejected in an earlier test.

The REGW *F* procedure is illustrated in Table 5.5-3. The critical values and decisions (significant, not significant, not significant by implication) are given in the fourth column of the table. The REGW *F* procedure resulted in rejecting three hypotheses for pairwise contrasts: $\mu_1 = \mu_4$, $\mu_1 = \mu_5$, and $\mu_2 = \mu_5$. All other hypotheses for pairwise contrasts are not rejected by implication because they are contained in sets of means—$\{\mu_1\ \mu_2\ \mu_3\}$, $\{\mu_2\ \mu_3\ \mu_4\}$, and $\{\mu_3\ \mu_4\ \mu_5\}$—that were not significant in earlier tests.

The REGW *F, FQ,* and *Q* procedures require critical values of α_r'' and *F* that are not available in the Studentized range and *F* tables. Values of α_r'' can be obtained by linear interpolation using the natural log of α_r''. For example, the following information from Appendix Table E.6 can be used to obtain the approximate critical value for $q_{.0303;\,3,\,40}$:

Critical Value From Appendix Table E.6	$\log_e \alpha_r''$
$q_{.01;\,3,\,40} = 4.37$	$\log_e .01 = -4.6052$
$q_{.0303;\,3,\,40} = \ ?$	$\log_e .0303 = -3.4966$
$q_{.05;\,3,\,40} = 3.44$	$\log_e .05 = -2.9957$

It is apparent that the critical value for $q_{.0303;\,3,\,40}$ is between 3.44 and 4.37. More precisely, the critical value is

$$\frac{\log_e .0303 - \log_e .05}{\log_e .01 - \log_e .05} = \frac{-3.4966 - (-2.9957)}{-4.6052 - (-2.9957)} = \frac{-0.5009}{-1.6094} = 0.3112$$

through the interval from 3.44 to 4.37. The critical value is

$$q_{.0303;\,3,\,40} \cong 3.44 + (0.3112)(4.37 - 3.44) = 3.44 + 0.2894 = 3.73$$

Microsoft's Excel FINV function can be used to obtain critical values for *F.* For example, the critical value for $F_{.0303,\,2,\,40}$ is obtained from the FINV function

FINV (probability,deg_freedom1,deg_freedom2)

FINV (.0303,2,40) = 3.8209

Monte Carlo studies (Einot & Gabriel, 1975; Ramsey, 1978, 1981; Seaman et al., 1991; Seaman, Levin, Serlin, & Franke, 1990) indicate that multiple comparison procedures that use an *F* statistic or an omnibus *F* statistic followed by a *q* statistic tend to be slightly more powerful than those that use an omnibus *q* statistic. These conclusions, however, are affected by the type of power and the pattern of means and sample sizes that were investigated. Unfortunately, the *F* statistic requires a considerable amount of computation, as can be seen from the example in Table 5.5-3.

The assumptions associated with the REGW *F* test are the same as those for the ANOVA *F* test discussed in Section 3.5. The assumptions associated with the REGW *FQ* test are the same as those for the ANOVA *F* test and Tukey's HSD test. If the sample sizes are unequal, the Tukey-Kramer test statistic can be used in place of the *q*REGW statistic. If the variances are unequal, one of the test statistics in the section on Procedures for Heterogeneous Variances can be used.

5.6 Testing all Contrasts Suggested by an Inspection of the Data

Scheffé's *S* Test

The fifth common research situation mentioned in Table 5.1-1 involves testing contrasts suggested by an inspection of the data when that inspection identifies one or more interesting nonpairwise contrasts. The procedure of choice for this situation is Scheffé's (1953) *S* test. The *S* test controls the familywise Type I error rate for the infinite number of contrasts that can be performed among $p \geq 3$ means. Scheffé's test is much less powerful than Tukey's HSD test, for example, and is recommended only when some nonpairwise contrasts are of interest. Scheffé's procedure uses the *F* sampling distribution and, like ANOVA, is robust with respect to nonnormality. The procedure also can be used when the sample sizes are unequal. Scheffé's test statistic, denoted by *FS,* is

$$FS = \frac{\hat{\psi}_i^2}{\hat{\sigma}_{\psi_i}^2} = \frac{\left(\sum_{j=1}^{p} c_j \overline{Y}_{\cdot j}\right)^2}{MS_{\text{error}} \sum_{j=1}^{p} \frac{c_j^2}{n_j}} = \frac{\left(c_1\overline{Y}_{\cdot 1} + c_2\overline{Y}_{\cdot 2} + \cdots + c_p\overline{Y}_{\cdot p}\right)^2}{MS_{\text{error}}\left(\frac{c_1^2}{n_1} + \frac{c_2^2}{n_2} + \cdots + \frac{c_p^2}{n_p}\right)}$$

A two-sided null hypothesis is rejected if *FS* exceeds or equals the critical value $\nu_1 F_{\alpha;\,\nu_1,\nu_2}$, where $F_{\alpha;\,\nu_1,\nu_2}$ is obtained from the *F* distribution in Appendix Table E.4, $\nu_1 = p - 1$, α denotes the familywise Type I error rate, and ν_2 is the degrees of freedom associated with MS_{error}.

Scheffé's procedure is always congruent with the omnibus ANOVA *F* test. If the omnibus *F* test is significant, at least one contrast among the means is significant according to Scheffé's test and vice versa. Scheffé's procedure is one of the most flexible data-snooping procedures available. But this flexibility comes at a price—low power. Hence, the procedure should be used only when the hypotheses of interest include a nonpairwise contrast.

Brown-Forsythe Test

If a researcher is interested in testing all contrasts, including nonpairwise contrasts that appear interesting from an inspection of the data, and if the population variances are heterogeneous, then the Brown-Forsythe (Brown & Forsythe, 1974a) procedure can be used. The procedure, which is a modification of Scheffé's procedure, uses the *F* sampling distribution and Welch's modified degrees of freedom. The test statistic, denoted by *FBF,* is

$$FBF = \frac{\hat{\psi}_i^2}{\hat{\sigma}_{\psi_i}^2} = \frac{\left(\sum_{j=1}^{p} c_j \overline{Y}_{\cdot j}\right)^2}{\sum_{j=1}^{p} \frac{c_j^2 \hat{\sigma}_j^2}{n_j}} = \frac{\left(c_1\overline{Y}_{\cdot 1} + c_2\overline{Y}_{\cdot 2} + \cdots + c_p\overline{Y}_{\cdot p}\right)^2}{\frac{c_1^2\hat{\sigma}_1^2}{n_1} + \frac{c_2^2\hat{\sigma}_2^2}{n_2} + \cdots + \frac{c_p^2\hat{\sigma}_p^2}{n_p}}$$

A two-sided null hypothesis is rejected if *FBF* exceeds or equals the critical value $\nu_1 F_{\alpha;\,\nu_1,\nu_2'}$, where $\nu_1 F_{\alpha;\,\nu_1,\nu_2'}$ is obtained from the *F* distribution in Appendix Table E.4, α denotes the familywise error rate, $\nu_1 = p - 1$, and ν_2' denotes Welch's modified degrees of freedom:

$$\nu_2' = \frac{\left(\frac{c_1^2\hat{\sigma}_1^2}{n_1} + \frac{c_2^2\hat{\sigma}_2^2}{n_2} + \cdots + \frac{c_p^2\hat{\sigma}_p^2}{n_p}\right)^2}{\frac{c_1^4\hat{\sigma}_1^4}{n_1^2(n_1-1)} + \frac{c_2^4\hat{\sigma}_2^4}{n_2^2(n_2-1)} + \cdots + \frac{c_p^4\hat{\sigma}_p^4}{n_p^2(n_p-1)}}$$

In practice, the single-step Brown-Forsythe procedure is usually preceded by an ANOVA *F* test of the omnibus null hypothesis. However, a preliminary ANOVA *F* test is not necessary because the Brown-Forsythe procedure controls the familywise Type I error rate. In fact, the test is very conservative. If a researcher is interested in only pairwise contrasts, one of the other procedures appropriate for heterogeneous variances, such as the Games-Howell test, should be used.

5.7 Other Multiple Comparison Procedures

Newman-Keuls and Duncan Tests

Table 5.1-2, discussed earlier, contains multiple comparison recommendations for the five common hypothesis-testing situations that occur in the behavioral sciences, health sciences, and education. The 17 recommended procedures control the per-contrast, familywise, or per-family Type I error rate for any complete or partial null hypothesis. In addition, each of the recommended procedures has one or more other virtues such as excellent power, ease of computation and interpretation, availability of confidence intervals, and robustness. Missing from the list are two nonrecommended tests: the Newman-Keuls test (Keuls, 1952; Newman, 1939) and Duncan's (1955) test. Both of these step-down procedures are used to test all pairwise contrasts among p means. They are competitors to the Tukey HSD, Fisher-Hayter, and REGW procedures described in Section 5.5. Researchers like the Newman-Keuls and Duncan procedures because of their excellent power. However, the Newman-Keuls procedure is not recommended because it fails to control the familywise Type I error rate when the family contains more than three means; Duncan's procedure fails to control the familywise Type I error rate when the family contains more than two means. Because of this serious shortcoming, I will say no more about these tests.

Peritz's Test

In 1970, Peritz introduced a step-down procedure that is a blend of the REGW and Newman-Keuls procedures. Peritz's procedure can be used with an F or a q statistic or a combination of an omnibus F statistic followed by a q statistic. The procedure controls the familywise Type I error and has been shown to have the highest per-pair power of all multiple comparison procedures investigated (Einot & Gabriel, 1975) and is among the highest in all-pairs power (Martin et al., 1989; Ramsey, 1981; Seaman et al., 1991). Seaman et al. (1990) have described two modified Peritz procedures that achieve a slight gain in power when $p > 4$. Unfortunately, the Peritz procedure and the two modifications are complex and are best performed with the aid of a computer. The interested reader can consult Begun and Gabriel (1981), Hochberg and Tamhane (1987), Kirk (1994), Ramsey (1981), or Toothaker (1991).

Controlling the False Discovery Rate

Throughout this chapter, I have emphasized the importance of controlling the familywise error rate for nonorthogonal contrasts. Testing a large number of contrasts can result in very low power for individual tests. Benjamini and Hochberg (1995) proposed controlling the **false discovery rate** (FDR) instead of the familywise error rate. The FDR is the expected proportion of contrasts falsely declared significant. Their idea was to make certain that the proportion of false discoveries relative to the total number of discoveries is kept small, say, no more than 5%. The false discovery rate can be defined as follows:

$$\text{False discovery rate } (\alpha_{\text{FDR}}) = \frac{\text{Number of contrasts falsely declared significant}}{\text{Number of contrasts declared significant}}$$

By controlling α_{FDR}, a researcher is less likely to make Type I errors than procedures that control the per-contrast error rate. And controlling α_{FDR} instead of α_{FW} provides more power to detect contrast that should be declared significant. When all null hypotheses are true, $\alpha_{\text{FDR}} = \alpha_{\text{FW}}$; when at least one null hypothesis is false, controlling α_{FDR} at, say, .05 means that α_{FW} will exceed .05. It follows that controlling the false discovery rate is not appropriate for all research situations. Control of α_{FDR} has been recommended for exploratory research and when the number of contrasts is extremely large (H. J. Keselman, Cribbie, & Holland, 1999).

Research on the merits of controlling the false discovery rate and on procedures for controlling the rate is in its infancy. The interested reader can consult Hemmelmann, Horn, Süsse, Vollandt, and Weiss (2005); Horn and Dunnett (2004); Korn, Troendle, McShane, and Simon (2004); Sakar (2002); and Somerville and Hemmelmann (2008).

5.8 Comparison of Multiple Comparison Procedures

Twenty-two multiple comparison procedures have been described in this chapter. The procedures and their salient characteristics are summarized in Table 5.8-1. The relative merits of various multiple comparison procedures have engendered much debate among statisticians in recent years. Each of the procedures in Table 5.8-1 has been recommended by one or more statisticians. The problem facing a researcher is to select the test statistic that provides the desired kind of protection against Type I errors and at the same time provides maximum power. The characteristics of the most frequently recommended procedures have been described in some detail along with pertinent references so that researchers can make informed choices.

5.9 Review Exercises

1. Terms to remember:

   a. contrast (comparison) (5.1)
   b. pairwise comparison (5.1)
   c. nonpairwise comparison (5.1)
   d. orthogonal contrast (5.1)
   e. a priori (planned) test (5.1)
   f. data snooping (5.1)
   g. a posteriori (unplanned) test (5.1)
   h. exploratory data analysis (5.1)

Table 5.8-1 ■ Comparison of Multiple Comparison Procedures

Test	Pairwise Contrasts Only	Pairwise or Nonpairwise Contrasts	Equal *ns*	Equal or Unequal *ns*	Homogeneous Variances	Heterogeneous Variances
A Priori Orthogonal Contrasts						
Student's t (5.2)*		X		X	X	
Student's t with Welch *df* (5.2)		X		X		X
$p - 1$ a Priori, Nonorthogonal Contrasts Involving a Control Group Mean						
Dunnett *tDN* (5.3)	X		X	X**	X	X**
C a Priori Nonorthogonal Contrasts						
Dunn (5.4)		X		X	X	
Dunn with Welch *df* (5.4)		X		X		X
Dunn-Šidák (5.4)		X		X	X	
Dunn-Šidák with Welch *df* (5.4)		X		X		X
Holm (5.4)		X		X	X	
Holm with Welch *df* (5.4)		X		X		X
All Pairwise Contrasts						
Fisher's LSD (5.1)	X			X	X	
Tukey's HSD (5.5)	X		X		X	
Tukey-Kramer (5.5)	X			X	X	
Dunnett's *T3* (5.5)	X			X		X
Dunnett's *C* (5.5)	X			X		X
Games-Howell (5.5)	X			X		X
Fisher-Hayter (5.5)	X			X	X	
REGW *F, FQ,* and *Q* (5.5)	X		X (*FQ, Q*)	X (*F*)	X	
Newman-Keuls (5.7)	X		X		X	
Duncan (5.7)	X		X		X	
Peritz *F, FQ,* and *Q* (5.7)	X		X (*FQ, Q*)	X (*F*)	X	
All Contrasts Including Nonpairwise Contrasts						
Scheffé (5.6)		X		X	X	
Brown-Forsythe (5.6)		X		X		X

*The numbers in parentheses denote the section where the test is described.

**With modification.

i. confirmatory data analysis (5.1)

j. family of contrasts (5.1)

k. per-contrast error rate (5.1)

l. familywise error rate (5.1)

m. per-family error rate (5.1)

n. experimentwise error rate (5.1)

o. overall power (5.1)

p. *P*-subset power (5.1)

q. any-pair power (5.1)

r. all-pairs power (5.1)

s. single-step procedure (5.1)

t. multiple-step procedure (5.1)

u. step-down procedure

v. coherence (5.1)

w. step-up procedure (5.1)

x. critical difference (5.4)

y. Dunn's multiple comparison procedure (5.4)

z. Bonferroni procedure (5.4)

aa. Fisher-Hayter test (5.5)

ab. false discovery rate (5.7)

2. [5.1] In order for $\psi_i = c_1\mu_1 + c_2\mu_2 + \cdots + c_p\mu_p$ to be a contrast, what conditions must the coefficients satisfy?

*3. [5.1] List the coefficients for the following contrasts.

*a. μ_1 versus μ_2

*b. μ_1 versus mean of μ_2 and μ_3

*c. Mean of μ_1 and μ_2 versus mean of μ_3 and μ_4

*d. μ_1 versus the weighted mean of μ_2 and μ_3, where μ_2 is weighted twice as much as μ_3

e. μ_1 versus mean of μ_2, μ_3, and μ_4

f. Mean of μ_1 and μ_2 versus mean of μ_3, μ_4, and μ_5

g. The weighted mean of μ_1 and μ_2 versus the weighted mean of μ_3 and μ_4, where μ_1 and μ_3 are weighted twice as much as μ_2 and μ_4

*4. [5.1] Which of the following meet the requirements for a contrast?

*a. $\mu_1 - \mu_2$

*b. $2\mu_1 - \mu_2 - \mu_3$

*c. $\mu_1 - \frac{1}{3}\mu_2 - \frac{1}{3}\mu_3$

*d. $\frac{3}{4}\mu_1 - \frac{1}{4}\mu_2 - \frac{1}{4}\mu_3 - \frac{1}{4}\mu_4$

e. $1\frac{1}{2}\mu_1 - \mu_2 - 1\frac{1}{2}\mu_3$

f. $\mu_1 - \frac{1}{4}\mu_2 - \frac{1}{4}\mu_3 - \frac{1}{4}\mu_4$

g. $\frac{1}{2}\mu_1 + \frac{1}{2}\mu_2 - \frac{1}{3}\mu_3 - \frac{1}{3}\mu_4 - \frac{1}{3}\mu_5$

*5. [5.1] Which contrasts in Exercise 4 satisfy $|c_1| + |c_2| + \cdots + |c_p| = 2$?

*6. [5.1] Indicate the number of pairwise comparisons that can be constructed for the following designs.

*a. CR-3 design

*b. CR-4 design

c. CR-5 design

d. CR-6 design

*7. [5.1] Which of the following sets of contrasts are orthogonal? Assume that the *n*s are equal.

*a. $\psi_1 = 1\mu_1 + (-1)\mu_2 + 0\mu_2$

$\psi_2 = 1\mu_1 + 0\mu_2 + (-1)\mu_3$

*b. $\psi_1 = 1\mu_1 + (-1)\mu_2 + 0\mu_3 + 0\mu_4$

$\psi_2 = 0\mu_1 + 0\mu_2 + 1\mu_3 + (-1)\mu_4$

*c. $\psi_1 = \frac{3}{4}\mu_1 + (-\frac{3}{4})\mu_2 + \frac{1}{4}\mu_3 + (-\frac{1}{4})\mu_4$

$\psi_2 = \frac{1}{2}\mu_1 + \frac{1}{2}\mu_2 + (-\frac{1}{2})\mu_3 + (-\frac{1}{2})\mu_4$

d. $\psi_1 = 1\mu_1 + (-1)\mu_2 + 0\mu_3 + 0\mu_4$

$\psi_2 = \frac{1}{2}\mu_1 + \frac{1}{2}\mu_2 + (-\frac{1}{2})\mu_3 + (-\frac{1}{2})\mu_4$

e. $\psi_1 = \frac{1}{2}\mu_1 + \frac{1}{2}\mu_2 + (-\frac{1}{2})\mu_3 + (-\frac{1}{2})\mu_4$

$\psi_2 = \frac{2}{3}\mu_1 + (-\frac{2}{3})\mu_2 + \frac{1}{3}\mu_3 + (-\frac{1}{3})\mu_4$

8. [5.1] Construct three sets of orthogonal contrasts among five means. Each set should contain four contrasts.

*9. [5.2] The religious dogmatism of members of four church denominations in a large midwestern city was investigated. A random sample of 30 members from each denomination took a paper-and-pencil test of dogmatism. The sample means were $\overline{Y}_{\cdot 1} = 64$, $\overline{Y}_{\cdot 2} = 73$, $\overline{Y}_{\cdot 3} = 61$, and $\overline{Y}_{\cdot 4} = 49$; $MSWG = 120$ and $\nu_2 = 4(30 - 1) = 116$. The researcher advanced the following a priori null hypotheses: $\mu_1 - \mu_2 = 0$, $\mu_3 - \mu_4 = 0$, and $(\mu_1 + \mu_2)/2 - (\mu_3 + \mu_4)/2 = 0$.

*a. Use a *t* statistic to test these null hypotheses; let $\alpha = .01$.

*b. Construct $100(1 - .01)\%$ confidence intervals for these a priori contrasts.

*c. [5.3] Compute the correlations among the contrasts.

*d. Assume that the sample variances are $\hat{\sigma}_1^2 = 62$, $\hat{\sigma}_2^2 = 73$, $\hat{\sigma}_3^2 = 80$, and $\hat{\sigma}_4^2 = 265$. Use Welch's t' statistic to test the null hypotheses.

10. [5.2] The effectiveness of three approaches to drug education in junior high school was investigated. The approaches were providing objective scientific information about the physiological and psychological effects of drug use, a_1; examining the psychology of drug use, a_2; and providing a control condition in which the chemical nature of various drugs was examined, a_3. Sixty-three students who did not use drugs were randomly assigned to the groups with the restriction that 21 were assigned to each group. At the conclusion of the educational program, the students evaluated its effectiveness; a high score signified effectiveness. The sample means were $\overline{Y}_{\cdot 1} = 25.8$, $\overline{Y}_{\cdot 2} = 26.7$, and $\overline{Y}_{\cdot 3} = 22.1$; $MSWG = 16.4$ and $\nu_2 = 3(21 - 1) = 60$. The researcher advanced the following a priori null hypotheses: $\mu_1 - \mu_2 = 0$ and $(\mu_1 + \mu_2)/2 - \mu_3 = 0$.

a. Use a *t* statistic to test these null hypotheses; let $\alpha = .05$.

b. Construct $100(1 - .05)\%$ confidence intervals for these a priori contrasts.

c. Compute the correlation between the contrasts.

d. Assume that the sample variances are $\hat{\sigma}_1^2 = 10.6$, $\hat{\sigma}_2^2 = 9.2$, and $\hat{\sigma}_3^2 = 29.4$. Use Welch's t' statistic to test the null hypotheses.

*11. [5.1] For the experiment in Exercise 9, what are the following error rates?

*a. Per contrast *b. Familywise *c. Per family

12. [5.1] For the experiment in Exercise 10, what are the following error rates?

a. Per contrast b. Familywise c. Per family

*13. [5.1] Suppose that 1000 experiments involving a CR-4 design have been performed, and in each experiment, hypotheses for all possible pairwise comparisons have been tested. Assume that 50 Type I errors are committed, and these occur in 35 of the 1000 experiments. Compute the following.

*a. Error rate per contrast *b. Error rate familywise

*c. Error rate per family

14. [5.1] Suppose that 1000 experiments involving a CR-5 design have been performed, and in each experiment, hypotheses for all possible pairwise comparisons have been tested. Assume that 80 Type I errors are committed, and these occur in 60 of the 1000 experiments. Compute the following.

a. Error rate per contrast b. Error rate familywise

c. Error rate per family

*15. [5.1] Compute α_{FW} for the following.

*a. Three a priori hypotheses involving orthogonal contrasts are each tested at $\alpha_{PC} = .01$.

*b. Four a priori hypotheses involving nonorthogonal contrasts are each tested at $\alpha_{PC} = .05$.

c. Four a priori hypotheses involving orthogonal contrasts are each tested at $\alpha_{PC} = .05$.

d. Five a priori hypotheses involving nonorthogonal contrasts are each tested at $\alpha_{PC} = .01$.

16. [5.1] For each of the following, indicate the recommended conceptual unit for error rate.

a. A priori orthogonal contrasts b. A priori nonorthogonal contrasts

c. A posteriori nonorthogonal contrasts

*17. [5.4] The effects of four doses of ethylene glycol on the reaction times of 20 chimpanzees were investigated. The animals were randomly assigned to one of four groups with five in each group. Those assigned to group a_1, the control group, received a placebo; those assigned to group a_2 received 0.2 fluid ounces (fl oz) of the drug; those assigned to group a_3 received 0.4 fl oz; and those assigned to group a_4 received 0.6 fl oz. The sample means were $\overline{Y}_{\cdot 1} = 0.28$ second, $\overline{Y}_{\cdot 2} = 0.29$ second, $\overline{Y}_{\cdot 3} = 0.31$ second, and $\overline{Y}_{\cdot 4} = .39$ second; $MSWG = 0.002$ and $\nu_2 = 4(5 - 1) = 16$. The researcher advanced a priori hypotheses about all pairwise comparisons among means.

*a. Use the Dunn-Šidák procedure to test these hypotheses by comparing $\hat{\psi}_i$ with $\hat{\psi}(tDS)$. Construct a table like Table 5.5-1; let $\alpha_{FW} = .05$.

*b. Construct 1 – 100(1 – .05)% confidence intervals for these a priori contrasts.

*c. Compute the correlations among the contrasts; assume that $c_{1j} = 1\ \ -1\ \ 0\ \ 0$, $c_{2j} = 1\ \ 0\ \ -1\ \ 0$, and so on.

*d. Compare the critical differences of the Dunn-Šidák statistic and Dunn statistic.

*e. Suppose that the researcher is interested only in the $p - 1 = 3$ contrasts that involve the control group. For this case, compare the critical differences of the Dunn and Dunn-Šidák procedures with Dunnett's statistic.

18. [5.4] The effects of information regarding a rape victim's past sexual behavior on perceived culpability were investigated. One hundred twenty-four college students were randomly assigned to one of four groups with 31 in each group. The students read specially written newspaper stories describing testimony at a trial. One newspaper account of the trial, condition a_1, indicated that the victim had an inactive sexual history. According to other accounts, the victim refused to discuss her past sexual experience, a_2; the judge prohibited testimony regarding past sexual history, a_3; and no mention of past sexual experience came up, a_4. The students rated the culpability of the victim on a 10-point scale; the higher the rating, the more culpable the victim. The sample means were $\overline{Y}_{\cdot 1} = 4.2$, $\overline{Y}_{\cdot 2} = 7.1$, $\overline{Y}_{\cdot 3} = 3.3$, and $\overline{Y}_{\cdot 4} = 4.5$; $MSWG = 14.08$ and $\nu_2 = 4(31 - 1) = 120$. The researcher advanced a priori hypotheses about all pairwise comparisons among means.

a. Use the Dunn-Šidák procedure to test these hypotheses by comparing $\hat{\psi}_i$ with $\hat{\psi}(tDS)$. Construct a table like Table 5.5-1; let $\alpha_{FW} = .05$.

b. Construct 1 – 100(1 – .05)% confidence intervals for these a priori contrasts.

c. Compute the correlations among the contrasts; assume that $c_{1j} = 1\ \ -1\ \ 0\ \ 0$, $c_{2j} = 1\ \ 0\ \ -1\ \ 1$, and so on.

d. Compare the critical difference of the Dunn-Šidák statistic with the Dunn statistic.

e. Suppose that the researcher is only interested in the $p - 1 = 3$ contrasts involving treatment level a_4. For this case, compare the critical differences of the Dunn and Dunn-Šidák statistics with Dunnett's statistic.

*19. [5.5] Exercise 17 described an experiment to evaluate the effects of four doses of ethylene glycol on the reaction time of chimpanzees.

*a. Use Tukey's procedure to test the omnibus null hypothesis $\mu_1 = \mu_2 = \mu_3 = \mu_4$. If this hypothesis is rejected, proceed to test all pairwise comparisons. Construct a table like Table 5.5-1; let $\alpha_{\text{FW}} = .05$.

*b. Construct $1 - 100(1 - .05)\%$ confidence intervals for all pairwise comparisons.

*c. Use the Holm test to evaluate all pairwise comparisons.

*d. Use the Fisher-Hayter test to evaluate all pairwise comparisons by comparing $\hat{\psi}_i$ with $\hat{\psi}(qFH)$.

*e. Use the REGW Q test to evaluate all pairwise comparisons.

*f. Rank the procedures in terms of apparent power.

20. [5.5] Exercise 18 described an experiment to evaluate the effects of information regarding a rape victim's past sexual behavior on perceived culpability.

a. Use Tukey's procedure to test the omnibus null hypothesis $\mu_1 = \mu_2 = \mu_3 = \mu_4$. If this hypothesis is rejected, proceed to test all pairwise comparisons. Construct a table like Table 5.5-1; let $\alpha_{\text{FW}} = .05$.

b. Construct $1 - 100(1 - .05)\%$ confidence intervals for all pairwise comparisons.

c. Use the Holm test to evaluate all pairwise comparisons.

d. Use the Fisher-Hayter test to evaluate all pairwise comparisons by comparing $\hat{\psi}_i$ with $\hat{\psi}(qFH)$.

e. Use the REGW Q test to evaluate all pairwise comparisons.

f. Rank the procedures in terms of apparent power.

*21. [5.6] Exercise 10 described an experiment to evaluate the effectiveness of three approaches to drug education in junior high school. Assume that the omnibus null hypothesis was rejected at the .05 level of significance.

*a. Use Scheffé's procedure to test the following null hypotheses:

$$H_0\colon \mu_1 - \mu_3 = 0 \qquad H_0\colon \mu_2 - \mu_3 = 0 \qquad H_0\colon (\mu_1 + \mu_2)/2 - \mu_3 = 0$$

Let $\alpha_{\text{FW}} = .05$.

*b. Suppose that the sample variances for this problem are $\hat{\sigma}_1^2 = 4.1$, $\hat{\sigma}_2^2 = 13.3$, and $\hat{\sigma}_3^2 = 31.8$. Use the Brown-Forsythe procedure to test the null hypotheses.

22. [5.6] The effects of simulator training involving synergistic 6-degrees-of-freedom platform motion on the acquisition of basic approach and landing skills of 63 undergraduate pilot trainees were investigated. The trainees were randomly divided into three groups. Those in group a_1 received 10 sorties with platform

motion in the Advanced Simulator for Pilot Training. Those in group a_2 also received 10 sorties but without motion. Trainees in group a_3, the control group, received the standard syllabus of preflight and flightline instructions. The dependent variable was instructor-pilot ratings of trainee performance in a T-37 aircraft. The sample means were $\overline{Y}_{\cdot 1} = 16.2$, $\overline{Y}_{\cdot 2} = 15.1$, and $\overline{Y}_{\cdot 3} = 11.4$; $MSWG = 39.94$ and $\nu_2 = 3(21 - 1) = 60$. Assume that the omnibus null hypothesis was rejected at the .05 level.

a. Use Scheffé's procedure to test the following null hypotheses:

$$H_0\colon \mu_1 - \mu_2 = 0 \qquad H_0\colon (\mu_1 + \mu_2)/2 - \mu_3 = 0$$

Let $\alpha_{FW} = .05$.

b. Suppose that the sample variances for this problem are $\hat{\sigma}_1^2 = 28.12$, $\hat{\sigma}_2^2 = 31.63$, and $\hat{\sigma}_3^2 = 60.07$. Use the Brown-Forsythe procedure to test the null hypotheses.