3 MULTIPLE COMPARISON TESTS

3.1 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. Other important uses are described in subsequent chapters. If the overall hypothesis of equality of means is rejected, an experimenter is still faced with the problem of deciding which of the means are not equal. Thus an overall F test is often merely the first step in analyzing a set of data. A significant F ratio indicates that something has happened in an experiment that has a small probability of happening by chance. The purpose of this chapter is to describe a variety of procedures for pinpointing what has happened. Specifically we will examine a number of test statistics for deciding which population means in an experiment are not equal. But first, several important concepts need to be defined.

CONTRAST DEFINED

A *contrast* or *comparison* among means is a difference among the means, with appropriate algebraic signs. We will use the symbols ψ_i and $\hat{\psi}_i$ to denote, respectively,

the ith contrast among population means and a sample estimate of the ith contrast. For example, $\hat{\psi}_i = \overline{Y}_{j\cdot} - \overline{Y}_{j'\cdot}$ is the ith contrast between sample means for treatment levels j and j'. If an experiment contains p equal to three treatment levels, contrasts involving two and three means may be of interest:

$$\begin{aligned} \hat{\psi}_1 &= \overline{Y}_{1\cdot} - \overline{Y}_{2\cdot} & \hat{\psi}_4 &= (\overline{Y}_{1\cdot} + \overline{Y}_{2\cdot})/2 - \overline{Y}_{3\cdot} \\ \hat{\psi}_2 &= \overline{Y}_{1\cdot} - \overline{Y}_{3\cdot} & \hat{\psi}_5 &= (\overline{Y}_{1\cdot} + \overline{Y}_{3\cdot})/2 - \overline{Y}_{2\cdot} \\ \hat{\psi}_3 &= \overline{Y}_{2\cdot} - \overline{Y}_{3\cdot} & \hat{\psi}_6 &= (\overline{Y}_{2\cdot} + \overline{Y}_{3\cdot})/2 - \overline{Y}_{1\cdot}. \end{aligned} \tag{3.1-1}$$

The contrasts on the right involve the average of two means versus a third mean.

More formally, a contrast or comparison among sample means is a linear combination of means with known coefficients, c_i's, such that (1) at least one coefficient is not equal to zero and (2) the coefficients sum to zero. That is,

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

where $c_j \neq 0$ for some j and $\sum_{j=1}^{p} c_j = 0$. If two of the coefficients are equal to 1 and -1 and all of the other coefficients are equal to zero, the contrast is called a *pairwise comparison;* otherwise it is a *nonpairwise comparison*. The contrasts in (3.1-1) can be expressed as linear combinations of the $\overline{Y}_{j\cdot}$'s by the appropriate choice of coefficients.

$\hat{\psi}_i =$	$c_1\overline{Y}_{1\cdot}$	$+ c_2\overline{Y}_{2\cdot}$	$+ c_3\overline{Y}_{3\cdot}$	
$\hat{\psi}_1 =$	$(1)\overline{Y}_{1\cdot}$	$+ (-1)\overline{Y}_{2\cdot}$	$+ (0)\overline{Y}_{3\cdot}$	$= \overline{Y}_{1\cdot} - \overline{Y}_{2\cdot}$
$\hat{\psi}_2 =$	$(1)\overline{Y}_{1\cdot}$	$+ (0)\overline{Y}_{2\cdot}$	$+ (-1)\overline{Y}_{3\cdot}$	$= \overline{Y}_{1\cdot} - \overline{Y}_{3\cdot}$
$\hat{\psi}_3 =$	$(0)\overline{Y}_{1\cdot}$	$+ (1)\overline{Y}_{2\cdot}$	$+ (-1)\overline{Y}_{3\cdot}$	$= \overline{Y}_{2\cdot} - \overline{Y}_{3\cdot}$
$\hat{\psi}_4 =$	$(\frac{1}{2})\overline{Y}_{1\cdot}$	$+ (\frac{1}{2})\overline{Y}_{2\cdot}$	$+ (-1)\overline{Y}_{3\cdot}$	$= (\overline{Y}_{1\cdot} + \overline{Y}_{2\cdot})/2 - \overline{Y}_{3\cdot}$
$\hat{\psi}_5 =$	$(\frac{1}{2})\overline{Y}_{1\cdot}$	$+ (-1)\overline{Y}_{2\cdot}$	$+ (\frac{1}{2})\overline{Y}_{3\cdot}$	$= (\overline{Y}_{1\cdot} + \overline{Y}_{3\cdot})/2 - \overline{Y}_{2\cdot}$
$\hat{\psi}_6 =$	$(-1)\overline{Y}_{1\cdot}$	$+ (\frac{1}{2})\overline{Y}_{2\cdot}$	$+ (\frac{1}{2})\overline{Y}_{3\cdot}$	$= (\overline{Y}_{2\cdot} + \overline{Y}_{3\cdot})/2 - \overline{Y}_{1\cdot}$

Note that for each contrast, $c_j \neq 0$ for some j and $\sum_{j=1}^{p} c_j = 0$. For convenience, coefficients of contrasts are often chosen so that the sum of their absolute values is equal to two. That is,

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

where $|c_j|$ indicates that the sign of c_j is always taken to be plus. All six contrasts above satisfy this property. For example, the sum of the absolute value of the coefficients for $\hat{\psi}_1$ and $\hat{\psi}_4$ are, respectively,

$$\begin{aligned} |1| + |-1| + |0| &= 1 + 1 + 0 = 2 \\ |\tfrac{1}{2}| + |\tfrac{1}{2}| + |-1| &= \tfrac{1}{2} + \tfrac{1}{2} + 1 = 2. \end{aligned}$$

The number of pairwise comparisons that can be defined for p means

is $p(p-1)/2$. Contrasts $\hat{\psi}_1$, $\hat{\psi}_2$, and $\hat{\psi}_3$ exhaust the $3(3-1)/2 = 3$ pairwise comparisons among the three means. The situation is quite different when we consider the number of nonpairwise comparisons that can be defined for $p \geq 3$ means. The number is infinite. Three nonpairwise comparisons for p equal to three means were given earlier: $\hat{\psi}_4$, $\hat{\psi}_5$, and $\hat{\psi}_6$. Other examples are

$$\hat{\psi}_7 = (1/3)\overline{Y}_{\cdot 1} + (2/3)\overline{Y}_{\cdot 2} + (-1)\overline{Y}_{\cdot 3}$$
$$\hat{\psi}_8 = (1/4)\overline{Y}_{\cdot 1} + (3/4)\overline{Y}_{\cdot 2} + (-1)\overline{Y}_{\cdot 3}$$
$$\hat{\psi}_9 = (1/5)\overline{Y}_{\cdot 1} + (4/5)\overline{Y}_{\cdot 2} + (-1)\overline{Y}_{\cdot 3}.$$

This particular pattern of coefficients can be extended indefinitely.

ORTHOGONAL CONTRASTS

As we have seen there is an infinite number of contrasts that can be formulated for $p \geq 3$ means. However, most of these contrasts can be expressed as linear combinations of other contrasts and as such involve redundant information. For example, $\hat{\psi}_3 = \overline{Y}_{\cdot 2} - \overline{Y}_{\cdot 3}$ defined earlier is equal to $\hat{\psi}_2 - \hat{\psi}_1$

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

and $\hat{\psi}_4 = (\overline{Y}_{\cdot 1} + \overline{Y}_{\cdot 2})/2 - \overline{Y}_{\cdot 3}$ is equal to $(1/2)\hat{\psi}_2 + (1/2)\hat{\psi}_3$

$$\hat{\psi}_4 = \overbrace{(\overline{Y}_{\cdot 1} - \overline{Y}_{\cdot 3})/2}^{(1/2)\hat{\psi}_2} + \overbrace{(\overline{Y}_{\cdot 2} - \overline{Y}_{\cdot 3})/2}^{(1/2)\hat{\psi}_3} = (\overline{Y}_{\cdot 1} + \overline{Y}_{\cdot 2})/2 - \overline{Y}_{\cdot 3}.$$

Sometimes an experimenter is interested in contrasts that are mutually nonredundant and uncorrelated. Such contrasts are called *orthogonal contrasts*. A simple rule exists for determining whether contrasts are orthogonal. Let $\hat{\psi}_i$ and $\hat{\psi}_{i'}$ denote two contrasts and c_{ij} and $c_{i'j}$ $(j = 1, \ldots, p)$ their respective coefficients. The 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 $\hat{\psi}_1 = (1)\overline{Y}_{\cdot 1} + (-1)\overline{Y}_{\cdot 2}$ and

* If the means are $NID(\mu_j, \sigma_\epsilon^2/n_j)$, orthogonality of contrasts is equivalent to statistical independence of the contrasts. The correlation between contrasts i and i' is given by

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

$\hat{\psi}_2 = (1)\overline{Y}_{\cdot 1} + (-1)\overline{Y}_{\cdot 3}$, and assume that the n_j's are equal. These contrasts are not orthogonal since

$$\begin{aligned} c_{1j} \quad \text{for } \hat{\psi}_1 &= c_{11}c_{12}c_{13} = 1 \quad -1 \quad 0 \\ c_{2j} \quad \text{for } \hat{\psi}_2 &= c_{21}c_{22}c_{23} = \underline{1 \quad 0 \quad -1} \\ \sum_{j=1}^{p} c_{1j}c_{2j} &= 1 + \ 0 + \ 0 = 1. \end{aligned}$$

However, contrasts $\hat{\psi}_1 = (1)\overline{Y}_{\cdot 1} + (-1)\overline{Y}_{\cdot 2}$ and $\hat{\psi}_4 = (\frac{1}{2})\overline{Y}_{\cdot 1} + (\frac{1}{2})\overline{Y}_{\cdot 2} + (-1)\overline{Y}_{\cdot 3}$ are orthogonal.

$$\begin{aligned} c_{1j} \quad \text{for } \hat{\psi}_1 &= c_{11}c_{12}c_{13} = 1 \quad -1 \quad 0 \\ c_{4j} \quad \text{for } \hat{\psi}_4 &= c_{41}c_{42}c_{43} = \underline{\tfrac{1}{2} \quad \tfrac{1}{2} \quad -1} \\ \sum_{j=1}^{p} c_{1j}c_{4j} &= \tfrac{1}{2} \quad -\tfrac{1}{2} \quad 0 = 0 \end{aligned}$$

The latter two contrasts, $\hat{\psi}_1$ and $\hat{\psi}_4$, exhaust one of the possible sets of orthogonal contrasts among three means. Two other sets of orthogonal contrasts are

$$\hat{\psi}_2 = (1)\overline{Y}_{\cdot 1} + (-1)\overline{Y}_{\cdot 3} \qquad \text{and} \qquad \hat{\psi}_5 = (\tfrac{1}{2})\overline{Y}_{\cdot 1} + (\tfrac{1}{2})\overline{Y}_{\cdot 3} + (-1)\overline{Y}_{\cdot 2}$$

since

$$\begin{aligned} c_{2j} \quad \text{for } \hat{\psi}_2 &= c_{21}c_{22}c_{23} = \ 1 \quad \ 0 \quad -1 \\ c_{5j} \quad \text{for } \hat{\psi}_5 &= c_{51}c_{52}c_{53} = \underline{\tfrac{1}{2} \quad -1 \quad \tfrac{1}{2}} \\ \sum_{j=1}^{p} c_{2j}c_{5j} &= \tfrac{1}{2} \quad \ 0 \quad -\tfrac{1}{2} = 0 \end{aligned}$$

and

$$\hat{\psi}_3 = (1)\overline{Y}_{\cdot 2} + (-1)\overline{Y}_{\cdot 3} \qquad \text{and} \qquad \hat{\psi}_6 = (\tfrac{1}{2})\overline{Y}_{\cdot 2} + (\tfrac{1}{2})\overline{Y}_{\cdot 3} + (-1)\overline{Y}_{\cdot 1}$$

since

$$\begin{aligned} c_{3j} \quad \text{for } \hat{\psi}_3 &= c_{31}c_{32}c_{33} = \ 0 \quad \ 1 \quad -1 \\ c_{6j} \quad \text{for } \hat{\psi}_6 &= c_{61}c_{62}c_{63} = \underline{-1 \quad \tfrac{1}{2} \quad \tfrac{1}{2}} \\ \sum_{j=1}^{p} c_{3j}c_{6j} &= \ 0 \quad \tfrac{1}{2} \quad -\tfrac{1}{2} = 0. \end{aligned}$$

For $p \geq 4$ means, an infinite number of sets of orthogonal contrasts exists. Table 3.1-1 gives seven sets involving four means.

A general principle can be stated. If an experiment contains p treatment levels, the number of orthogonal contrasts in any set is equal to $p - 1$. Furthermore, an orthogonal set provides a basis for constructing all other contrasts. That is, all other contrasts can be expressed as linear combinations of those in an orthogonal set. For example, consider Set 1 in Table 3.1-1. The six contrasts in Sets 2 and 3 can be expressed as linear combinations of the three orthogonal contrasts in Set 1 as follows:

$$\text{Set 2}\quad \begin{cases} \hat{\psi}_4 = (\hat{\psi}_1 - \hat{\psi}_2)/2 + \hat{\psi}_3 = \overline{Y}_{\cdot 1} - \overline{Y}_{\cdot 3} \\ \hat{\psi}_5 = (-\hat{\psi}_1 + \hat{\psi}_2)/2 + \hat{\psi}_3 = \overline{Y}_{\cdot 2} - \overline{Y}_{\cdot 4} \\ \hat{\psi}_6 = (\hat{\psi}_1 + \hat{\psi}_2)/2 = (\overline{Y}_{\cdot 1} + \overline{Y}_{\cdot 3})/2 - (\overline{Y}_{\cdot 2} + \overline{Y}_{\cdot 4})/2 \end{cases}$$

$$\text{Set 3}\quad \begin{cases} \hat{\psi}_7 = (\hat{\psi}_1 + \hat{\psi}_2)/2 + \hat{\psi}_3 = \overline{Y}_{\cdot 1} - \overline{Y}_{\cdot 4} \\ \hat{\psi}_8 = (-\hat{\psi}_1 - \hat{\psi}_2)/2 + \hat{\psi}_3 = \overline{Y}_{\cdot 2} - \overline{Y}_{\cdot 3} \\ \hat{\psi}_9 = (\hat{\psi}_1 - \hat{\psi}_2)/2 = (\overline{Y}_{\cdot 1} + \overline{Y}_{\cdot 4})/2 - (\overline{Y}_{\cdot 2} + \overline{Y}_{\cdot 3})/2. \end{cases}$$

TABLE 3.1-1 Sets of Orthogonal Contrasts Among Four Means (Numbers in the table are the coefficients of the contrasts)

Set	$c_1\overline{Y}_{\cdot 1} + c_2\overline{Y}_{\cdot 2} + c_3\overline{Y}_{\cdot 3} + c_4\overline{Y}_{\cdot 4}$				*Contrast*
1	1	−1	0	0	$\hat{\psi}_1 = \overline{Y}_{\cdot 1} - \overline{Y}_{\cdot 2}$
	0	0	1	−1	$\hat{\psi}_2 = \overline{Y}_{\cdot 3} - \overline{Y}_{\cdot 4}$
	1/2	1/2	−1/2	−1/2	$\hat{\psi}_3 = (\overline{Y}_{\cdot 1} + \overline{Y}_{\cdot 2})/2 - (\overline{Y}_{\cdot 3} + \overline{Y}_{\cdot 4})/2$
2	1	0	−1	0	$\hat{\psi}_4 = \overline{Y}_{\cdot 1} - \overline{Y}_{\cdot 3}$
	0	1	0	−1	$\hat{\psi}_5 = \overline{Y}_{\cdot 2} - \overline{Y}_{\cdot 4}$
	1/2	−1/2	1/2	−1/2	$\hat{\psi}_6 = (\overline{Y}_{\cdot 1} + \overline{Y}_{\cdot 3})/2 - (\overline{Y}_{\cdot 2} + \overline{Y}_{\cdot 4})/2$
3	1	0	0	−1	$\hat{\psi}_7 = \overline{Y}_{\cdot 1} - \overline{Y}_{\cdot 4}$
	0	1	−1	0	$\hat{\psi}_8 = \overline{Y}_{\cdot 2} - \overline{Y}_{\cdot 3}$
	1/2	−1/2	−1/2	1/2	$\hat{\psi}_9 = (\overline{Y}_{\cdot 1} + \overline{Y}_{\cdot 4})/2 - (\overline{Y}_{\cdot 2} + \overline{Y}_{\cdot 3})/2$
4	1	−1	0	0	$\hat{\psi}_{10} = \overline{Y}_{\cdot 1} - \overline{Y}_{\cdot 2}$
	1/2	1/2	−1	0	$\hat{\psi}_{11} = (\overline{Y}_{\cdot 1} + \overline{Y}_{\cdot 2})/2 - \overline{Y}_{\cdot 3}$
	1/3	1/3	1/3	−1	$\hat{\psi}_{12} = (\overline{Y}_{\cdot 1} + \overline{Y}_{\cdot 2} + \overline{Y}_{\cdot 3})/3 - \overline{Y}_{\cdot 4}$
5	1/2	1/2	−1/2	−1/2	$\hat{\psi}_{13} = (\overline{Y}_{\cdot 1} + \overline{Y}_{\cdot 2})/2 - (\overline{Y}_{\cdot 3} + \overline{Y}_{\cdot 4})/2$
	1/2	−1/2	−1/2	1/2	$\hat{\psi}_{14} = (\overline{Y}_{\cdot 1} + \overline{Y}_{\cdot 4})/2 - (\overline{Y}_{\cdot 2} + \overline{Y}_{\cdot 3})/2$
	−1/2	1/2	−1/2	1/2	$\hat{\psi}_{15} = (\overline{Y}_{\cdot 2} + \overline{Y}_{\cdot 4})/2 - (\overline{Y}_{\cdot 1} + \overline{Y}_{\cdot 3})/2$
6	1/2	1/2	−1/2	−1/2	$\hat{\psi}_{16} = (\overline{Y}_{\cdot 1} + \overline{Y}_{\cdot 2})/2 - (\overline{Y}_{\cdot 3} + \overline{Y}_{\cdot 4})/2$
	2/3	−2/3	1/3	−1/3	$\hat{\psi}_{17} = (2\overline{Y}_{\cdot 1} + \overline{Y}_{\cdot 3})/3 - (2\overline{Y}_{\cdot 2} + \overline{Y}_{\cdot 4})/3$
	−1/3	1/3	2/3	−2/3	$\hat{\psi}_{18} = (\overline{Y}_{\cdot 2} + 2\overline{Y}_{\cdot 3})/3 - (\overline{Y}_{\cdot 1} + 2\overline{Y}_{\cdot 4})/3$
7	1/2	1/2	−1/2	−1/2	$\hat{\psi}_{19} = (\overline{Y}_{\cdot 1} + \overline{Y}_{\cdot 2})/2 - (\overline{Y}_{\cdot 3} + \overline{Y}_{\cdot 4})/2$
	3/4	−3/4	1/4	−1/4	$\hat{\psi}_{20} = (3\overline{Y}_{\cdot 1} + \overline{Y}_{\cdot 3})/4 - (3\overline{Y}_{\cdot 2} + \overline{Y}_{\cdot 4})/4$
	−1/4	1/4	3/4	−3/4	$\hat{\psi}_{21} = (\overline{Y}_{\cdot 2} + 3\overline{Y}_{\cdot 3})/4 - (\overline{Y}_{\cdot 1} + 3\overline{Y}_{\cdot 4})/4$

As we have seen, there are always $p - 1$ nonredundant questions that can be answered from the data in an experiment. However, an experimenter may not be interested in all of the $p - 1$ questions. For example, an experimenter may want to test the hypotheses that $\mu_1 - \mu_2 = 0$ and $\mu_3 - \mu_4 = 0$ but not that $(\mu_1 + \mu_2)/2 - (\mu_3 + \mu_4)/2 = 0$. The latter hypothesis may have no meaning in terms of the objectives of the experiment. Also, not all interesting questions involve orthogonal contrasts. In an experiment with four treatment levels, each of the six pairwise comparisons among means may be associated with a question that the experimenter seeks to answer.

In summary, the analysis of variance provides an overall test of the hypothesis that $\mu_1 = \mu_2 = \cdots = \mu_p$ or $\alpha_j = 0$ for all j. This test is equivalent to a simultaneous test of the hypothesis that all possible contrasts among means are equal to zero. The degrees of freedom for *MSBG* in a completely randomized design is $p - 1$, which is also the number of orthogonal contrasts that can be constructed from p means. If an overall F test is significant, an experimenter can be certain that some set of orthogonal contrasts contains at least one significant contrast among the means. The contrast or contrasts that are significant may or may not be ones that are of interest to the experimenter. An F test in ANOVA is an overall test that indicates whether something has happened. It remains for an experimenter to carry out follow-up tests to determine what has happened. The following sections describe procedures for carrying out tests of (1) planned orthogonal contrasts, (2) planned nonorthogonal contrasts, and (3) unplanned nonorthogonal contrasts.

3.2 A PRIORI ORTHOGONAL CONTRASTS

In planning an experiment, one often has a specific set of hypotheses that the experiment is designed to test. Tests involving these hypotheses are referred to as *a priori* or *planned* tests. This situation may be contrasted with another in which an investigator believes that a treatment affects the dependent variable and the experiment is designed to accept or reject this notion. If an overall F test indicates that at least one contrast is not equal to zero, interest turns to determining which contrast or contrasts among means is significant. Tests that are used for *data snooping*—that is, for evaluating a subset of all possible contrasts following a significant overall test—are referred to as *a posteriori, unplanned,* or *post hoc* tests.

A PRIORI ORTHOGONAL TESTS USING A t STATISTIC

Hypotheses of the form

$$H_0: \psi_1 = 0$$
$$H_0: \psi_2 = 0$$
$$\cdots$$

$$H_0: \quad \psi_i \quad = 0$$
$$\cdots$$
$$H_0: \quad \psi_{p-1} = 0$$

where $\psi_i = c_1\mu_1 + c_2\mu_2 + \cdots + c_p\mu_p$ and the $p - 1$ contrasts are mutually orthogonal and a priori, can be tested using a t statistic. It is not necessary to perform an overall test of significance prior to testing planned orthogonal contrasts. An overall test using, say, an F statistic simply answers the question, "Did anything happen in the experiment?" If a specific set of hypotheses for orthogonal contrasts has been advanced, an experimenter is not interested in answering this general question. Rather, one is interested in answering a limited number—$p - 1$ or fewer—specific questions from the data. In such cases, it is recommended that each hypothesis be evaluated at α level of significance. The rationale for this recommendation is discussed later.

A t statistic for testing the hypothesis $H_0: \quad \psi_i = 0$ is given by

$$\textbf{(3.2-1)} \qquad t = \frac{\hat{\psi}_i}{\hat{\sigma}_{\psi_i}} = \frac{\sum_{j=1}^{p} c_j \overline{Y}_j}{\sqrt{MS_{\text{error}} \sum_{j=1}^{p} \frac{c_j^2}{n_j}}} = \frac{c_1\overline{Y}_1 + c_2\overline{Y}_2 + \cdots + c_p\overline{Y}_p}{\sqrt{MS_{\text{error}}\left(\frac{c_1^2}{n_1} + \frac{c_2^2}{n_2} + \cdots + \frac{c_p^2}{n_p}\right)}}$$

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 a completely randomized design,

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

with $p(n - 1)$ degrees of freedom. If the sample sizes are not equal, the formula for unequal n_j's in Section 2.2 is used to compute $MSWG$. Under the assumptions that

1. The observations are drawn from normally distributed populations or the sample n_j's are fairly large,
2. The observations are random samples from the populations,
3. The null hypothesis is true, and
4. The variances of the $j = 1, \ldots, p$ populations are equal to σ_ϵ^2,

the t statistic (3.2-1) is distributed as Student's t with $p(n - 1)$ degrees of freedom. The critical values of t that cut off the upper α and $\alpha/2$ regions of Student's t distribution for ν degrees of freedom are given in Appendix Table E.4 and denoted by $t_{\alpha,\nu}$ and $t_{\alpha/2,\nu}$, respectively. The null hypothesis for a two-tailed test is rejected if the absolute value of t, $|t|$, exceeds $t_{\alpha/2,\nu}$. For $H_0: \quad \mu_j - \mu_{j'} \leq 0$, t must exceed $t_{\alpha,\nu}$; for $H_0: \quad \mu_j - \mu_{j'} \geq 0$, $-t$ must be less than $-t_{\alpha,\nu}$.

Multiple t statistics use the same error mean square 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.

COMPUTATIONAL EXAMPLE OF A PRIORI ORTHOGONAL TESTS USING A t STATISTIC

The use of multiple t statistics to test hypotheses about a priori orthogonal contrasts will be illustrated for an experiment in which five qualitative treatment levels have been randomly assigned to 50 subjects. Ten subjects receive each treatment level. Assume that the five treatment means are

$$\overline{Y}_1 = 36.7, \quad \overline{Y}_2 = 48.7, \quad \overline{Y}_3 = 43.4, \quad \overline{Y}_4 = 47.2, \quad \text{and} \quad \overline{Y}_5 = 40.3.$$

Also assume that the treatment populations are approximately normally distributed and that the variances are homogeneous. The layout for this experiment corresponds to that for a completely randomized design; hence, $MSWG$ is the appropriate estimator of the common population error variance. The estimate is 28.8 with degrees of freedom equal to $p(n - 1) = 5(10 - 1) = 45$. The experiment has been designed to test the hypotheses in Table 3.2-1. The .05 level of significance is adopted for each test. The reader can easily verify from the coefficients in Table 3.2-1 that the $p - 1 = 4$ contrasts are mutually orthogonal. The t statistics are

$$t = \frac{\hat{\psi}_1}{\hat{\sigma}_{\psi_1}} = \frac{(1)48.7 + (-1)43.4}{\sqrt{28.8\left[\frac{(1)^2}{10} + \frac{(-1)^2}{10}\right]}} = \frac{5.300}{2.400} = 2.21$$

$$t = \frac{\hat{\psi}_2}{\hat{\sigma}_{\psi_2}} = \frac{(1)47.2 + (-1)40.3}{\sqrt{28.8\left[\frac{(1)^2}{10} + \frac{(-1)^2}{10}\right]}} = \frac{6.900}{2.400} = 2.88$$

$$t = \frac{\hat{\psi}_3}{\hat{\sigma}_{\psi_3}} = \frac{(\tfrac{1}{2})48.7 + (\tfrac{1}{2})43.4 + (-\tfrac{1}{2})47.2 + (-\tfrac{1}{2})40.3}{\sqrt{28.8\left[\frac{(\frac{1}{2})^2}{10} + \frac{(\frac{1}{2})^2}{10} + \frac{(-\frac{1}{2})^2}{10} + \frac{(-\frac{1}{2})^2}{10}\right]}} = \frac{2.300}{1.697} = 1.36$$

$$t = \frac{\hat{\psi}_4}{\hat{\sigma}_{\psi_4}} = \frac{(1)36.7 + (-\tfrac{1}{4})48.7 + (-\tfrac{1}{4})43.4 + (-\tfrac{1}{4})47.2 + (-\tfrac{1}{4})40.3}{\sqrt{28.8\left[\frac{(1)^2}{10} + \frac{(-\frac{1}{4})^2}{10} + \frac{(-\frac{1}{4})^2}{10} + \frac{(-\frac{1}{4})^2}{10} + \frac{(-\frac{1}{4})^2}{10}\right]}}$$

$$= \frac{-8.200}{1.897} = -4.32.$$

The critical value required to reject the null hypothesis is, according to Appendix Table E.4, $t_{.05/2,45} = 2.02$. Thus, the null hypothesis can be rejected for contrasts ψ_1, ψ_2, and ψ_4.

TABLE 3.2-1 Statistical Hypotheses and Associated Orthogonal Coefficients

Contrast	*Coefficients of Contrast*					*Hypotheses*
	c_1	c_2	c_3	c_4	c_5	
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	1/2	1/2	−1/2	−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	−1/4	−1/4	−1/4	−1/4	H_0: $\mu_1 - (\mu_2 + \mu_3 + \mu_4 + \mu_5)/4 = 0$ H_1: $\mu_1 - (\mu_2 + \mu_3 + \mu_4 + \mu_5)/4 \neq 0$

A PRIORI ORTHOGONAL TESTS USING AN *F* STATISTIC

We saw in Section 2.1 that t^2 is identical to F with 1 and ν_2 degrees of freedom. It is sometimes convenient to perform a priori orthogonal tests using an F statistic rather than a t statistic. The formula for an F statistic is

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

with 1 and ν_2 degrees of freedom. If *MSWG* is used as an estimator of the common population error variance, ν_2 is equal to $p(n - 1)$ when the n's are equal and $N - p$ when they are unequal.

CONFIDENCE INTERVALS FOR A PRIORI ORTHOGONAL CONTRASTS

Emphasis in this book is placed on significance tests as opposed to confidence intervals. This emphasis is in line with contemporary practice in the behavioral sciences and education. Many mathematical statisticians, however, prefer confidence interval procedures, and there is much merit in their position. Most of the hypothesis testing procedures described in this chapter can also be used to establish $100(1 - \alpha)\%$ confidence intervals for population contrasts.

A $100(1 - \alpha)\%$ confidence interval for an a priori orthogonal contrast is given by

$$\hat{\psi}_i - \hat{\psi}(c) \leq \psi_i \leq \hat{\psi}_i + \hat{\psi}(c)$$

$$\text{where} \quad \hat{\psi}(c) = t_{\alpha/2,\nu}\sqrt{MS_{\text{error}}\sum_{j=1}^{p}\frac{c_j^2}{n_j}}$$

$$\hat{\psi}_i = \sum_{j=1}^{p} c_j\overline{Y}_j$$

$$\psi_i = \sum_{j=1}^{p} c_j\mu_j.$$

To illustrate, a 95% confidence interval for contrast ψ_1 in Table 3.2-1, which involves the difference between μ_2 and μ_3, is given by

$$(48.7 - 43.4) - 4.8 \le \psi_1 \le (48.7 - 43.4) + 4.8$$
$$5.3 - 4.8 \le \psi_1 \le 5.3 + 4.8$$
$$0.5 \le \psi_1 \le 10.1$$

$$\text{where} \quad \hat{\psi}(c) = 2.02\sqrt{28.8\left(\frac{(1)^2}{10} + \frac{(-1)^2}{10}\right)}$$
$$= 2.02(2.4) = 4.8.$$

The two boundaries

$$(\overline{Y}_2 - \overline{Y}_3) - \hat{\psi}(c) = 0.5 \qquad \text{and} \qquad (\overline{Y}_2 - \overline{Y}_3) + \hat{\psi}(c) = 10.1$$

of the confidence interval are called the $100(1 - .05)\% = 95\%$ confidence limits. This confidence interval indicates values of $\mu_2 - \mu_3$ that are consistent with the observed sample means. After the samples have been obtained and point estimates of μ_2 and μ_3 have been substituted in the confidence interval, it is incorrect to say that the probability is .95 that the difference between the population means lies between 0.5 and 10.1. Once the interval has been computed, the difference $\mu_2 - \mu_3$ either is or is not in the interval. This point can be clarified by realizing that there are many 95% confidence intervals over all possible random samples. Some of these confidence intervals will include the true difference and others will not. If one confidence interval is sampled at random, the probability is .95 that it will include the true difference.

If, as in the present example, a confidence interval does not include zero, the hypothesis that the contrast equals zero can also be rejected. Confidence interval procedures permit an experimenter to reach the same kind of decision as tests of significance. In addition, confidence interval procedures permit an experimenter to consider simultaneously all possible null hypotheses, not just that the contrast equals zero. If the hypothesized value of the contrast lies outside the $100(1 - \alpha)\%$ confidence interval, the null hypothesis can be rejected at α level of significance. The size of the confidence interval also provides information concerning the error variation associated with an estimate and, hence, the strength of an inference. The preference of many mathematical statisticians for confidence interval procedures over significance tests is understandable since both procedures involve the same assumptions, but confidence interval procedures provide an experimenter with more information.

ROBUST PROCEDURES FOR A PRIORI ORTHOGONAL CONTRASTS

Significance tests and confidence intervals using a t statistic involve the assumptions that the populations are approximately normally distributed and the variances are homogeneous. The t statistic like the F statistic is robust with respect to violation of both assumptions provided that the number of observations in the samples is equal. However, Boik (1975) and Kohr and Games (1977) have shown that when the sample sizes or the absolute values of the contrast coefficients, $|c_j|$, are unequal (for example, ½, ½, 1), the t statistic is not robust to heterogeneity of variance. In the discussion that follows we will consider some test statistics that are robust under these conditions. Pairwise comparisons will be considered first.

When the population variances are unequal, the pooled estimator, MS_{error}, in the t statistic can be replaced by the individual variance estimators for populations j and j' as follows

$$t' = \frac{\hat{\psi}_i}{\hat{\sigma}_{\psi_i}} = \frac{c_j\overline{Y}_j + c_{j'}\overline{Y}_{j'}}{\sqrt{\frac{\hat{\sigma}_j^2}{n_j} + \frac{\hat{\sigma}_{j'}^2}{n_{j'}}}}. \tag{3.2-2}$$

The earliest attempts to determine the sampling distribution of t' were made by Behrens (1929) and enlarged upon by Fisher (1935). No exact solution for this problem exists. Three approximate solutions have been proposed: (1) Cochran and Cox (1957, 100); (2) Dixon and Massey (1957, 123), Satterthwaite (1946), and Smith (1936); and (3) Welch (1947).* In general, there is close agreement among these approximate solutions; accordingly only those of Cochran and Cox and Welch will be described.

Cochran and Cox's procedure uses the t' statistic defined in (3.2-2). The critical value of t' is given by

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

where $t_{\alpha/2,\nu_j}$ and $t_{\alpha/2,\nu_{j'}}$ are the tabled t values at $\alpha/2$ level of significance for $\nu_j = n_j - 1$ and $\nu_{j'} = n_{j'} - 1$ degrees of freedom, respectively. The critical value for t' will always be between the ordinary t values for ν_j and $\nu_{j'}$ degrees of freedom. For a one-tailed test, values of t_{α,ν_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 (1947, 1949) procedure for pairwise contrasts also uses the t' statistic defined in (3.2-2). Tables of the distribution of t' have been prepared by Aspin

* A different approach to the problem has been suggested by Johnson (1978). He used the Cornish-Fisher expansion to derive a corrected form of the t statistic. Properties of the data are used to adjust t so that the resulting statistic is approximately distributed as Student's t distribution.

(1949). An approximation to the critical value of t' can be obtained from Student's t distribution with degrees of freedom equal to

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

This procedure provides a relatively powerful test that is robust under all conditions that have been investigated (Scheffé, 1970; Wang, 1971).

For nonpairwise contrasts, Welch's (1947) t' test statistic is

$$t' = \frac{\hat{\psi}_i}{\hat{\sigma}_{\psi_i}} = \frac{c_1\overline{Y}_1 + c_2\overline{Y}_2 + \cdots + c_p\overline{Y}_p}{\sqrt{\frac{c_1^2\hat{\sigma}_1^2}{n_1} + \frac{c_2^2\hat{\sigma}_2^2}{n_2} + \cdots + \frac{c_p^2\hat{\sigma}_p^2}{n_p}}}$$

with degrees of freedom equal to

$$\nu' = \frac{v^2}{w}$$

where $$v = \frac{c_1^2\hat{\sigma}_1^2}{n_1} + \frac{c_2^2\hat{\sigma}_2^2}{n_2} + \cdots + \frac{c_p^2\sigma_p^2}{n_p}$$

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

According to Kohr and Games (1977), this 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. Under these conditions, Welch's procedures appear to be good choices for evaluating planned orthogonal contrasts.

Earlier it was recommended that each hypothesis for planned orthogonal contrasts be evaluated at α level of significance. In the sections that follow we will examine the rationale for this recommendation and describe procedures for evaluating planned nonorthogonal and unplanned nonorthogonal contrasts.

3.3 CONCEPTUAL UNIT FOR ERROR RATE

If an experimenter tests C independent contrasts, each at α level of significance, the probability of making one or more type I errors is

$$\text{Prob. of one or more type I errors} = 1 - (1 - \alpha)^C$$

which is approximately equal to $C\alpha$ for small values of α.

The rationale underlying $1 - (1 - \alpha)^C$ is as follows. We know that if a contrast is tested at α level of significance, the probability of not making a type I error is $1 - \alpha$. If C independent contrasts are each tested at α level of significance, the probability of not making a type I error for the first, and the second, . . . , and the Cth null hypothesis is, according to the multiplication rule for independent events,

$$\overbrace{(1 - \alpha)(1 - \alpha) \cdots (1 - \alpha)}^{C} = (1 - \alpha)^C.$$

This is the probability of retaining all C null hypotheses when they are true. The probability of not retaining all null hypotheses when they are true is

$$\text{Prob. of one or more type I errors} = 1 - (\text{Prob. of retaining all null hypotheses when they are true})$$
$$= 1 - (1 - \alpha)^C.$$

As the number of independent tests increases, so does the probability of obtaining spuriously significant results. For example, if α is equal to .05 and an experimenter tests three, five, or ten independent contrasts, the probability of one or more type I errors is, respectively,

$$1 - (1 - .05)^3 = .14$$
$$1 - (1 - .05)^5 = .23$$
$$1 - (1 - .05)^{10} = .40.$$

Hence, if enough t statistics are computed, each at α level of significance, an experimenter will probably reject one or more null hypotheses even though they are all true. This problem can be minimized by restricting the use of multiple t tests to a priori orthogonal contrasts. The corresponding probability of making one or more type I errors for nonindependent contrasts is difficult to compute, although it can be shown to be less than or equal to $1 - (1 - \alpha)^C$.* The basic question can be raised, "Should the probability of committing a type I error be set at α for each test or should the probability of one or more errors be set at α or less for some larger conceptual unit such as the collection of tests?" This question, which has been debated extensively in the literature, does not have a simple answer. Answers have varied, depending on the nature of the contrasts that are of interest to the experimenter. As we will see, the most relevant considerations appear to be whether the contrasts are orthogonal and whether they are planned in advance.

The meaning of the term significance level is unambiguous for experiments with two treatment levels, but not so when the experiment contains three or more

* Harter (1957) and Pearson and Hartley (1942, 1943) describe procedures for computing this probability.

treatment levels. The ambiguity arises because a significance level or *error rate* (*ER*) can be defined for a number of different conceptual units: the individual contrast, family of contrasts, and experiment. These conceptual units are described next.

ERROR RATE PER CONTRAST

The error rate per contrast, α_{PC}, is the probability that a contrast will be falsely declared significant. We can think of this in more concrete terms. Suppose that hypotheses for many many contrasts are tested and somehow we are able to count the number of erroneous conclusions, then

$$ER \text{ per contrast } (\alpha_{PC}) = \frac{\text{Number of contrasts falsely declared significant}}{\text{Number of contrasts}}.$$

When a t statistic is used to test a priori orthogonal contrasts each at α level of significance, the conceptual unit for error rate is the individual contrast. Controlling the error rate per contrast, however, allows the probability of making one or more type I errors for the set of contrasts to increase as the number of tests increases.

ERROR RATE PER EXPERIMENT AND EXPERIMENTWISE

An alternative strategy to controlling error rate per contrast is to adopt the experiment as the conceptual unit for error rate. Consider performing many experiments in which hypotheses for $C = 5$ contrasts are tested in each experiment and again assume that we are able to count the number of erroneous conclusions. Two error rates can be defined for this situation:

$$ER \text{ per experiment } (\alpha_{PE}) = \frac{\text{Number of contrasts falsely declared significant}}{\text{Number of experiments}}$$

and

$$ER \text{ experimentwise } (\alpha_{EW}) = \frac{\text{Number of experiments with at least one contrast falsely declared significant}}{\text{Number of experiments}}.$$

The first error rate is the long-run average number of erroneous statements made per experiment. The error rate experimentwise is the probability that one or more erroneous statements will be made in an experiment and is less conservative than the error rate per experiment. The experimentwise error rate is a probability, whereas the error rate per experiment is not a probability, but, rather, the expected number of errors per experiment. The use of the experimentwise error rate is based on the premise that it is as serious to make one erroneous statement in an experiment as it is to make, say, five such statements. The two error rates are numerically almost identical for small values of α. As a result, Miller (1966, 10) observes that a choice between them

is essentially a matter of taste. The relationship between error rate experimentwise and error rate per contrast for C orthogonal contrasts is

$$\alpha_{EW} = 1 - (1 - \alpha_{PC})^C.$$

For nonorthogonal contrasts, the relationship is

$$\alpha_{EW} \leq 1 - (1 - \alpha_{PC})^C.$$

The error rate experimentwise cannot exceed the error rate per experiment, that is, $\alpha_{EW} \leq \alpha_{PE}$.

When an F statistic is used to test the overall null hypothesis

$$H_0: \quad \mu_1 = \mu_2 = \cdots = \mu_p$$

in a single-treatment analysis of variance design at α level of significance, the experiment is the conceptual unit for error rate. Suppose that the overall null hypothesis is rejected. At this point, the interest usually shifts to determining which pairwise comparisons, if any, among means are significant. It is generally recommended that the error rate for the collection of a posteriori nonorthogonal follow-up tests equal that for the overall F test. As we will see, a variety of data snooping statistics have been developed for this purpose.

ERROR RATE PER FAMILY AND FAMILYWISE

In multitreatment ANOVA designs another conceptual unit for error rate can be adopted—the family. A family of contrasts consists of all contrasts of interest that are associated with a particular treatment or interaction. A factorial experiment with two treatments, for example, contains the three families of contrasts: one associated with treatment A, a second associated with treatment B, and a third associated with the AB interaction. Two new error rates can be defined for such experiments:

$$ER \text{ per family } (\alpha_{PF}) = \frac{\text{Number of contrasts falsely declared significant}}{\text{Number of families}}$$

$$ER \text{ familywise } (\alpha_{FW}) = \frac{\text{Number of families with at least one contrast falsely declared significant}}{\text{Number of families}}.$$

In multitreatment ANOVA designs it is customary to use an F statistic to test each of the overall null hypotheses at α level of significance. In other words, contemporary practice favors the family rather than the experiment as the conceptual unit for error rate. This is reasonable considering the special nature of the treatments and interactions in factorial and hierarchical designs—they are planned in advance and mutually orthogonal.* Consequently, instead of designing an experiment with, say,

* The orthogonality of sources of variations in ANOVA is discussed by Box, Hunter, and Hunter (1978, Ch. 6).

two treatments and one interaction, one could choose to design three separate experiments to test the three overall null hypotheses. Then one would have three separate experiments instead of one experiment with three a priori orthogonal families. But in either case, an F statistic would be used to test each of the three overall null hypotheses at α level of significance. The approach of combining the three families into one experiment has the advantage of providing more degrees of freedom for estimating the experimental error.*

Three conceptual units for error rate have been described: the contrast, family, and experiment. They are all identical for an experiment involving one contrast. The error rates become more disparate as the number of families and contrasts within families increases.

WHAT IS THE CORRECT CONCEPTUAL UNIT FOR ERROR RATE

The relative merits of making the contrast or some larger unit, such as the family or experiment, the conceptual unit for error rate has been extensively debated: Duncan (1955), McHugh and Ellis (1955), Ryan (1959, 1960, 1962), and Wilson (1962). The answer to the question, "What is the correct conceptual unit for error rate?" depends, as we have seen, on the nature of the contrasts. The following discussion summarizes recommendations for three categories of contrasts: (1) a priori and orthogonal, (2) a priori and nonorthogonal, and (3) a posteriori and nonorthogonal.

The interpretation of significance level for an experiment involving only one contrast is unambiguous. However, the situation becomes more complicated when an experiment involves several contrasts. If orthogonal contrasts have been planned in advance, contemporary practice favors the contrast as the conceptual unit for error rate. We noted in Section 3.1 that testing planned orthogonal contrasts is equivalent to partitioning the data so that each test involves nonredundant, independent pieces of information. The same can be said for treatments and interactions in multitreatment ANOVA designs. In this case, contemporary practice favors the family of contrasts associated with a treatment or interaction as the conceptual unit for error rate.

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 conceptual unit for error rate, either the family or, in the case of single-treatment experiments, the experiment. Mathematical statisticians have developed a wide variety of test statistics for controlling error rate at or less than α for various collections of tests. For example, test statistics have been developed for controlling error rate at or less than α for (1) the comparison of a control-group mean with $p - 1$ experimental-group means, (2) any set of C contrasts among means, (3) all $p(p - 1)/2$ pairwise comparisons among means, and (4) all possible contrasts among means. Test statistics in the first two categories are well suited to evaluating

* As noted in Section 3.2, when the same error term is used in a series of tests, the tests are not statistically independent even though the sources being tested are independent.

a priori nonorthogonal contrasts; those in the latter two categories are more often used for a posteriori nonorthogonal contrasts, although exceptions occur. Hard and fast rules are difficult to formulate because, as we will see, the various test statistics differ markedly in power. The problem facing an experimenter is to select the test statistic that provides the desired kind of protection and the maximum power. In general, test statistics that were designed for testing a select, limited number of contrasts are more powerful than those designed to test all pairwise comparisons or all possible contrasts. Hence, when possible, it is to the experimenter's advantage to specify a select, limited number of contrasts in advance. The strategy of using more powerful test statistics for planned tests and less powerful statistics for data snooping is discussed in Section 4.4.

3.4 A PRIORI NONORTHOGONAL CONTRASTS

A researcher, in planning an experiment, often has a specific set of C hypotheses that the experiment is designed to test. Often the associated contrasts are not orthogonal as in comparing a control-group mean with $p - 1$ experimental group means, comparing each mean with every other mean, or making, say, $C = p + 2$ tests among p means. If the contrasts are planned in advance and relatively limited in number, one of the test statistics in this section can be used. If the contrasts are a posteriori or if the number is relatively large, one of the test statistics in Section 3.5 should be considered. Often several test statistics will provide acceptable protection against making one or more type I errors. In such cases, the experimenter is encouraged to compute the critical difference necessary to reject the null hypotheses for the alternative test procedures and use the one that gives the smallest critical difference (Miller, 1966, 18). As we will see, this is one of the more useful techniques for choosing among alternative test procedures.

DUNN'S MULTIPLE COMPARISON PROCEDURE

Dunn's procedure is appropriate for testing hypotheses concerning C planned contrasts among means. The originator of the procedure is unknown. Dunn (1961) has examined the properties of the procedure in detail and has prepared tables that facilitate its use. Consequently it is referred to as *Dunn's multiple comparison procedure*. Some writers refer to it as the *Bonferroni t procedure*, since it is based on the Bonferroni or Boole inequality.

Dunn (1961) used the Bonferroni inequality in showing that the error rate experimentwise* cannot exceed the sum of C per contrast error rates, that is,

* The term experimentwise is used here because the following computational example involves a type CR-p design. If a multitreatment design were used, we would control the error rate familywise.

$$\alpha_{EW} \leq \sum_{i=1}^{C} \alpha_{PC_i}$$

where α_{PC_i} is the per contrast error rate for the ith contrast. Thus, if each of C contrasts is tested at α/C level of significance, the error rate experimentwise cannot exceed α. The procedure basically consists of splitting up α among a set of C planned contrasts. For example, if we want to test C equal to 2 contrasts and we want the error rate experimentwise to be less than or equal to .05, each contrast should be tested at $.05/2 = .025$ level of significance. Then, $\alpha_{EW} \leq .025 + .025 = .05$. It is not necessary to perform an overall test of significance prior to testing the planned contrasts.

Dunn developed the procedure using Student's t distribution and the t statistic described in Section 3.2, but it is applicable to other test statistics. The t statistic will be denoted by tD when it is used with Dunn's procedure. The tD statistic is

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

We will denote the critical value for this tD statistic by $tD_{\alpha/2;C,\nu}$. The ith null hypothesis H_0: $\psi_i = 0$ is rejected if $|tD| \geq tD_{\alpha/2;C,\nu}$, where $tD_{\alpha/2;C,\nu}$ is the two-tailed critical value of tD in Appendix Table E.16*, C is the number of planned contrasts among p means, and ν is the degrees of freedom for MS_{error}. The critical values in Table E.16 were obtained from Student's t distribution; it can be shown for a two-tailed test that

$$tD_{\alpha/2;C,\nu} = t_{(\alpha/2)/C,\nu}.$$

For example, $tD_{.05/2;5,20} = t_{(.05/2)/5,20} = 2.845$, where $t_{(.05/2)/5,20}$ cuts off the upper $(.05/2)/5 = .005$ portion of Student's t distribution.

Let's assume that an experimenter is interested in testing the following hypotheses involving a priori nonorthogonal contrasts:

H_0: $\psi_1 = \mu_1 - \mu_2 = 0$ $\quad$ H_0: $\psi_3 = \mu_4 - \mu_5 = 0$

H_0: $\psi_2 = \mu_1 - \mu_3 = 0$ $\quad$ H_0: $\psi_4 = (\mu_1 + \mu_2 + \mu_3)/3 - (\mu_4 + \mu_5)/2 = 0.$

Let the experimentwise error rate equal .01. Assume that the sample means are $\overline{Y}_1 = 36.7$, $\overline{Y}_2 = 48.7$, $\overline{Y}_3 = 43.4$, $\overline{Y}_4 = 47.2$, $\overline{Y}_5 = 40.3$ and that $MSWG = 28.8$ with $p(n-1) = 5(10-1) = 45$ degrees of freedom. These same data were used to illustrate the t statistic in Section 3.2. The values of the test statistics are

$$tD = \frac{\hat{\psi}_1}{\hat{\sigma}_{\psi_1}} = \frac{(1)36.7 + (-1)48.7}{\sqrt{28.8\left[\frac{(1)^2}{10} + \frac{(-1)^2}{10}\right]}} = \frac{-12.000}{2.400} = -5.00$$

* More complete tables are provided by Dayton and Schafer (1973).

$$tD = \frac{\hat{\psi}_2}{\hat{\sigma}_{\psi_2}} = \frac{(1)36.7 + (-1)43.4}{\sqrt{28.8\left[\frac{(1)^2}{10} + \frac{(-1)^2}{10}\right]}} = \frac{-6.700}{2.400} = -2.79$$

$$tD = \frac{\hat{\psi}_3}{\hat{\sigma}_{\psi_3}} = \frac{(1)47.2 + (-1)40.3}{\sqrt{28.8\left[\frac{(1)^2}{10} + \frac{(-1)^2}{10}\right]}} = \frac{6.900}{2.400} = 2.88$$

$$tD = \frac{\hat{\psi}_4}{\hat{\sigma}_{\psi_4}} = \frac{(1/3)36.7 + (1/3)48.7 + (1/3)43.4 + (-1/2)47.2 + (-1/2)40.3}{\sqrt{28.8\left[\frac{(1/3)^2}{10} + \frac{(1/3)^2}{10} + \frac{(1/3)^2}{10} + \frac{(-1/2)^2}{10} + \frac{(-1/2)^2}{10}\right]}}$$

$$= \frac{-0.817}{1.549} = -0.53.$$

The critical value of tD in Table E.16 is $tD_{.01/2;4,45} \cong 3.21$; hence, only the null hypothesis for ψ_1 can be rejected.

If an experimenter had advanced one-tailed hypotheses, the required value of $tD_{.01;4,45}$ could not be obtained from Table E.16. An approximate value of $tD_{\alpha;C,\nu}$ that cuts off the upper α/C proportion of Student's t distribution for ν degrees of freedom can be determined from the standard normal distribution by

$$tD_{\alpha;C,\nu} \cong z_{\alpha/C} + \frac{z_{\alpha/C}^3 + z_{\alpha/C}}{4(\nu - 2)}$$

where $z_{\alpha/C}$ is obtained from Appendix Table E.3 (Peiser, 1943). For our example,

$$tD_{.01;4,45} = t_{.01/4,45} \cong 2.81 + \frac{(2.81)^3 + 2.81}{4(45 - 2)} = 2.96$$

where 2.81 is the value of z that cuts off the upper $.01/4 = .0025$ proportion of the standard normal distribution.

If Dunn's multiple comparison procedure is used to test hypotheses concerning all pairwise comparisons among means and the sample n_j's are equal, it is computationally more convenient to compute the *critical difference* $\hat{\psi}(D)$ that a comparison must exceed in order to be declared significant. This difference is given by

$$\hat{\psi}(D) = tD_{\alpha/2;C,\nu}\sqrt{2MS_{\text{error}}/n}.$$

Suppose that an experimenter wants to evaluate all pairwise comparisons among the five means given earlier. One can compute $5(5 - 1)/2 = 10$ such comparisons. The values of the comparisons are given in Table 3.4-1, where the means have been ordered from the smallest to the largest. The critical difference that a pairwise comparison must exceed is

$$\hat{\psi}(D) = tD_{.01/2;10,45}\sqrt{2(28.8)/10}$$
$$= 3.520(2.40) = 8.45.$$

TABLE 3.4-1 Absolute Value of Differences Among Means ($MS_{error} = 28.8$, $p = 5$, and $n = 10$)

	$\overline{Y}_1$	$\overline{Y}_5$	$\overline{Y}_3$	$\overline{Y}_4$	$\overline{Y}_2$
$\overline{Y}_1 = 36.7$	—	3.6	6.7	10.5*	12.0*
$\overline{Y}_5 = 40.3$		—	3.1	6.9	8.4
$\overline{Y}_3 = 43.4$			—	3.8	5.3
$\overline{Y}_4 = 47.2$				—	1.5
$\overline{Y}_2 = 48.7$					—

*$p < .01$

Two differences, the ones that are starred in Table 3.4-1, exceed 8.45. Thus, the null hypotheses for these differences can be rejected.

In both of the preceding examples, the level of significance α was divided evenly among the contrasts by using Dunn's table (Table E.16). This procedure of dividing α evenly among C contrasts is appropriate if an experimenter considers the consequences of making a type I error to be equally serious for all contrasts. If this is not true, α can be allocated unequally among the C contrasts in a manner that reflects the experimenter's a priori concern for type I and II errors. Let's assume that the .05 level of significance is adopted for a collection of C equal to five a priori contrasts. The use of Dunn's table in Appendix E amounts to testing each of the five contrasts at α_i, where $\alpha_i = \alpha/C = .05/5 = .01$. Suppose that the consequences of making a type I error are not equally serious for all contrasts. If this is true, the experimenter can allocate α_i unequally among the contrasts in a manner that reflects concern about type I errors, so long as the sum of α_i for $i = 1, \ldots, C$ contrasts is equal to α, the significance level selected for the collection of tests. For example, the five values of α_i could be $\alpha_1 = .02$, $\alpha_2 = .01$, $\alpha_3 = .01$, $\alpha_4 = .005$, and $\alpha_5 = .005$. The error rate for the collection of the five tests is equal to $.02 + .01 + .01 + .005 + .005 = .05$, which is the same value that would have been obtained if α were divided equally among the five tests.

Dunn's procedure can also be used to establish C simultaneous $100(1 - \alpha)\%$ confidence intervals for a collection of population contrasts. The degree of one's confidence that all C contrasts are simultaneously in their respective confidence intervals is greater than or equal to $1 - \sum_{i=1}^{C} \alpha_{PC_i}$. The confidence interval is given by

$$\hat{\psi}_i - \hat{\psi}(D) \leq \psi_i \leq \hat{\psi}_i + \hat{\psi}(D)$$

where $$\hat{\psi}(D) = tD_{\alpha/2;C,\nu}\sqrt{MS_{error}\sum_{j=1}^{p}\frac{c_j^2}{n_j}}$$

$$\hat{\psi}_i = \sum_{j=1}^{p} c_j\overline{Y}_j$$

$$\psi_i = \sum_{j=1}^{p} c_j\mu_j.$$

ŠIDÁK'S MODIFICATION OF DUNN'S PROCEDURE

Significance levels and confidence coefficients for Dunn's procedure are approximate since they are based on the additive Bonferroni inequality. For C nonorthogonal contrasts, we saw that

$$\alpha_{EW} \leq \sum_{i=1}^{C} \alpha_{PC_i}.$$

It is evident from this additive inequality that Dunn's procedure provides an upper bound to the error rate experimentwise; that is, the error rate experimentwise cannot exceed $\sum_{i=1}^{C} \alpha_{PC_i}$. For small values of α_{EW}, the approximation is excellent. However, an even better approximation to the upper bound for α_{EW} is provided by a multiplicative inequality proved by Šidák (1967). According to Šidák, the experimentwise error rate for nonorthogonal contrasts is always less than or equal to $1 - (1 - \alpha)^C$.* The following relations can be shown to hold for nonorthogonal contrasts

$$\alpha_{EW} \leq 1 - (1 - \alpha_{PC})^C < \sum_{i=1}^{C} \alpha_{PC_i}.$$

It follows that instead of testing each contrast at α_{EW}/C level of significance to control α_{EW} as in Dunn's procedure, each contrast can be tested at $1 - (1 - \alpha_{EW})^{1/C}$ level of significance. Differences between the critical values for the two procedures are negligible for $\alpha_{EW} < .01$. Suppose an experimenter plans to test five nonorthogonal contrasts and wants to control the probability of making one or more errors at or less than .05. If $\alpha_{EW} = .05$ is split evenly among the five contrasts, use of the additive and multiplicative inequalities would result in testing each contrast at, respectively,

Additive inequality $\quad \dfrac{\alpha_{EW}}{C} = \dfrac{.05}{5} = .01$

Multiplicative inequality $\quad 1 - (1 - \alpha_{EW})^{1/C} = 1 - (1 - .05)^{1/5} = .010206.$

Use of .010206 instead of .01 requires a slightly smaller critical value and critical difference. Hence the multiplicative inequality leads to a more powerful test and a narrower confidence interval than the Bonferroni additive inequality. Note, however, that it is relatively easy to allocate α unequally among C contrasts using the additive inequality; this is not the case for the multiplicative inequality. In subsequent discussions we will refer to the procedure based on the additive inequality as Dunn's procedure and that based on the multiplicative inequality as the Dunn-Šidák procedure.

Games (1977) developed a table of critical values for the t statistic based on the multiplicative inequality $\alpha_{EW} \leq 1 - (1 - \alpha_{PC})^C$; these values are given in Appendix Table E.17. The t statistic will be denoted by *tDS* when it is used with the Dunn-Šidák procedure. The *tDS* statistic is

* Dunn (1958) proved this earlier for $p = 2$, 3, or any p with a special set of variances and covariances (variance-covariance matrix).

$$tDS = \frac{\hat{\psi}_i}{\hat{\sigma}_{\psi_i}} = \frac{\sum_{j=1}^{p} c_j \overline{Y}_j}{\sqrt{MS_{\text{error}} \sum_{j=1}^{p} \frac{c_j^2}{n_j}}}.$$

We will denote the value that t must exceed for C contrasts by $tDS_{\alpha/2;C,\nu}$. The critical difference that a contrast must exceed is denoted by $\hat{\psi}(DS)$ and is equal to

$$\hat{\psi}(DS) = tDS_{\alpha/2;C,\nu} \sqrt{MS_{\text{error}} \sum_{j=1}^{p} \frac{c_j^2}{n_j}}.$$

If an experimenter tests five nonorthogonal contrasts at $\alpha_{EW} = .05$ and there are 60 degrees of freedom for experimental error, the critical values for Dunn's and Šidák's procedures are, respectively,

$$tD_{\alpha/2;5,60} = 2.660$$
$$tDS_{\alpha/2;5,60} = 2.653.$$

Both procedures control the probability of making one or more type I errors at or less than α; however, the Dunn-Šidák procedure is always slightly more powerful.

Because these procedures are not restricted to orthogonal contrasts but are applicable to any number of planned contrasts, the reader may wonder why they have not replaced the multiple t statistic described in Section 3.2. The answer is to be found in a comparison of the length of the confidence interval for the procedures. For $C \geq 2$, a confidence interval based on the Dunn or Dunn-Šidák procedures is always longer than the corresponding interval based on the multiple t statistic procedure. The advantage of being able to make all planned contrasts, not just those that are orthogonal, is gained at the expense of an increase in the probability of making type II errors.

Dunn (1961) compared her procedure with a posteriori procedures developed by Tukey (1953) and Scheffé (1953). Games (1977) made a similar comparison for the Dunn-Šidák and Scheffé procedures. The Tukey and Scheffé procedures, which are described in Section 3.5, were developed for exploring interesting contrasts suggested by an inspection of the data. Dunn and Games have shown that when there are many means in an experiment, and the number of contrasts that an experimenter wants to evaluate is relatively small, the Dunn and Dunn-Šidák procedures produce shorter confidence intervals than either of the a posteriori procedures.

On the other hand, if p is small and C is relatively large, the a posteriori procedures lead to shorter confidence intervals. This results, in part, because the length of the confidence interval for the Dunn and Dunn-Šidák procedures depends on C, the number of contrasts among p means, whereas with both a posteriori procedures the length depends on p, the number of means. For planned contrasts, an experimenter can always determine in advance which multiple comparison procedure will lead to the smallest critical difference for a contrast or shortest confidence interval.

DUNNETT'S TEST FOR CONTRASTS INVOLVING A CONTROL MEAN

The object of many experiments is to compare $p - 1$ treatment means with a control-group mean. Dunnett (1955) has developed a multiple comparison procedure for this purpose, that is, for testing $p - 1$ a priori null hypotheses of the form

$$\begin{aligned} H_0: &\quad \psi_1 = \mu_1 - \mu_p = 0 \\ H_0: &\quad \psi_2 = \mu_2 - \mu_p = 0 \\ &\cdots \\ H_0: &\quad \psi_{p-1} = \mu_{p-1} - \mu_p = 0. \end{aligned}$$

More specifically, Dunnett's procedure is applicable to a special set of $p - 1$ a priori nonorthogonal comparisons—those in which the correlation between any two contrasts is 0.5. This occurs when $p - 1$ means are compared with a control-group mean. To illustrate, consider the following contrasts where $n = 10$ and $\overline{Y}_4$ is the control-group mean.

	$\overline{Y}_1$	$\overline{Y}_2$	$\overline{Y}_3$	$\overline{Y}_4$
$\hat{\psi}_1 =$	1	0	0	$-1 = \overline{Y}_1 - \overline{Y}_4$
$\hat{\psi}_2 =$	0	1	0	$-1 = \overline{Y}_2 - \overline{Y}_4$
$\hat{\psi}_3 =$	0	0	1	$-1 = \overline{Y}_3 - \overline{Y}_4$

The correlation between the ith and the i'th contrasts is given by

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

The correlations for the ith and the i'th contrasts are

$$\begin{aligned} \rho_{12} &= (1/10)/\sqrt{(2/10)(2/10)} = 0.5 \\ \rho_{13} &= (1/10)/\sqrt{(2/10)(2/10)} = 0.5 \\ \rho_{23} &= (1/10)/\sqrt{(2/10)(2/10)} = 0.5. \end{aligned}$$

For such contrasts, Dunnett's procedure controls the probability of falsely rejecting one or more null hypotheses at α_{EW}. It is not necessary to perform an overall test of significance prior to testing the planned comparisons.

The test statistic for Dunnett's procedure is the t statistic described in Section 3.2. We will denote this statistic by tD' when it is used with Dunnett's procedure.

$$tD' = \frac{\hat{\psi}_i}{\hat{\sigma}_{\psi_i}} = \frac{c_j\overline{Y}_j + c_{j'}\overline{Y}_{j'}}{\sqrt{MS_{\text{error}}\left(\frac{c_j^2}{n_j} + \frac{c_{j'}^2}{n_{j'}}\right)}}$$

If the sample n's are equal, the statistic simplifies to

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

The two-tailed critical value for this t statistic is denoted by $tD'_{\alpha/2;p,\nu}$, where p is the number of means, including the control mean, and ν is the degrees of freedom for MS_{error}. The one-tailed critical value is denoted by $tD'_{\alpha;p,\nu}$. The critical values are given in Appendix Table E.9.

Dunnett's multiple comparison procedure will be illustrated for the data in Table 3.4-1; let $\overline{Y}_1$ be the control mean. Four pairwise comparisons involving the control mean can be made among these five means. The critical difference $\hat{\psi}(D')$ that a comparison must exceed for a two-tailed test at $\alpha_{EW} = .01$ is given by

$$\begin{aligned}\hat{\psi}(D') &= tD'_{.01;5,45}\sqrt{2MS_{\text{error}}/n} \\ &= 3.167\sqrt{2(28.8)/10} \\ &= 7.60.\end{aligned}$$

From an examination of Table 3.4-1, it is apparent that the absolute value of two comparisons involving $\overline{Y}_1$ exceeds 7.60; these are $|\overline{Y}_1 - \overline{Y}_2| = 12$ and $|\overline{Y}_1 - \overline{Y}_4| = 10.5$. Hence the null hypotheses for these comparisons can be rejected.

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

$$\hat{\psi}_i - \hat{\psi}(D') \leq \psi_i \leq \hat{\psi}_i + \hat{\psi}(D')$$

where $\hat{\psi}(D') = tD'_{\alpha/2;p,\nu}\sqrt{2MS_{\text{error}}/n}$

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

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

Dunnett's procedure and the Dunn and Dunn-Šidák procedures can be used to evaluate a priori nonorthogonal contrasts. However, Dunnett's procedure is restricted to $p - 1$ comparisons where the correlation between each pair of contrasts equals 0.5. For this application, Dunnett's procedure is exact whereas the Dunn and Dunn-Šidák procedures only provide an upper bound to the experimentwise error rate. Insight concerning the relative efficiency of the three multiple comparison procedures can be obtained from

$$\text{Relative efficiency} = \frac{(\text{Critical difference for more efficient test})^2}{(\text{Critical difference for less efficient test})^2} \times 100.$$

Suppose an experimenter plans to compare four treatment means with a control mean. Let $MS_{\text{error}} = 28.8$, $n = 10$, $\nu = 45$, and $\alpha_{EW} = .01$. The critical differences for the Dunn, Dunn-Šidák, and Dunnett procedures are, respectively,

$$\begin{aligned}\hat{\psi}(D) &= tD_{.01/2;4,45}\sqrt{2MS_{\text{error}}/n} = 3.20667(2.40) = 7.696 \\ \hat{\psi}(DS) &= tDS_{.01/2;4,45}\sqrt{2MS_{\text{error}}/n} = 3.20167(2.40) = 7.684 \\ \hat{\psi}(D') &= tD'_{.01/2;5,45}\sqrt{2MS_{\text{error}}/n} = 3.16667(2.40) = 7.600.\end{aligned}$$

The efficiency of the Dunn and Dunn-Šidák procedures relative to Dunnett's procedure is, respectively,

$$\text{Relative efficiency} = \frac{[\hat{\psi}(D')]^2}{[\hat{\psi}(D)]^2} \times 100 = \frac{(7.600)^2}{(7.696)^2} \times 100 = 97.5\%$$

$$\text{Relative efficiency} = \frac{[\hat{\psi}(D')]^2}{[\hat{\psi}(DS)]^2} \times 100 = \frac{(7.600)^2}{(7.684)^2} \times 100 = 97.8\%.$$

For comparing $p - 1$ treatment means with a control mean, Dunnett's procedure is the method of choice.

A measure of the relative length of the confidence interval for any two procedures is given by

$$\text{Relative length} = \frac{\text{Length of shorter confidence interval}}{\text{Length of longer confidence interval}} \times 100.$$

According to the following computations, the length of the confidence interval for Dunnett's procedure is approximately 99% as long as that for the Dunn and Dunn-Šidák procedures.

$$\text{Relative length} = \frac{\hat{\psi}(D')}{\hat{\psi}(D)} \times 100 = \frac{(7.600)}{(7.696)} \times 100 = 98.8\%$$

$$\text{Relative length} = \frac{\hat{\psi}(D')}{\hat{\psi}(DS)} \times 100 = \frac{(7.600)}{(7.684)} \times 100 = 98.9\%$$

Dunnett (1964) has described modifications of his procedure that can be used when the variance of the control group is not equal to the variance of the $p - 1$ treatment groups.

3.5 A POSTERIORI NONORTHOGONAL CONTRASTS

Many experiments are designed to determine if any treatment effects are present. If a test of significance leads to rejection of the overall null hypothesis, attention is directed to exploring the data in order to find the source of the effects. A number of procedures has been developed for data snooping. Most are appropriate for evaluating all pairwise comparisons among means; one, Scheffé's S method, can be used to evaluate all contrasts among means. Unless indicated otherwise, all of the procedures assume that

1. The observations are drawn from normally distributed populations or the sample n_j's are fairly large.
2. The observations are random samples from the populations.
3. The null hypothesis is true.
4. The variances of the $j = 1, \ldots, p$ populations are equal to σ_ϵ^2.

As we will see, the procedures differ markedly in the protection they offer against making type I and II errors.

LEAST SIGNIFICANT DIFFERENCE TEST

The first a posteriori test described here is also one of the oldest. In 1935, Fisher (1949, 56–58) described a multiple comparison procedure called the *least significant difference* (*LSD*) test. This test consists of first performing a test of the overall null hypothesis that $\mu_1 = \mu_2 = \cdots = \mu_p$ by means of an F statistic. If the overall null hypothesis is rejected, multiple t statistics are used to evaluate all pairwise comparisons among means. If the overall F statistic is not significant, no further tests are performed. The error rate experimentwise is equal to α for the overall F test. However, if subsequent tests are performed, the conceptual unit for error rate is the individual comparison. Thus the *LSD* test is not consistent with respect to the error rate protection at the two stages of the test. This procedure has been widely used in research but is not generally recommended by statisticians—for an exception, see Carmer and Swanson (1973).

If the F statistic is significant, the least significant difference between two means, $\hat{\psi}(LSD)$, is

$$\hat{\psi}(LSD) = t_{\alpha/2,\nu}\sqrt{\frac{2MS_{\text{error}}}{n}}$$

where $t_{\alpha/2,\nu}$ is the upper $\alpha/2$ percentage point from Student's t distribution (Appendix Table E.4) and ν is the degrees of freedom associated with MS_{error}, the denominator of the F statistic. If the absolute value of a comparison $|\hat{\psi}| = |\overline{Y}_j - \overline{Y}_{j'}|$ exceeds $\hat{\psi}(LSD)$, the comparison is declared significant. This procedure is convenient if the n's are equal because $\hat{\psi}(LSD)$ need only be computed once for any set of comparisons. If the n's are not equal, multiple comparisons among means can be made using the formula (3.2-1) for the t statistic in Section 3.2.

The use of the *LSD* test can lead to an anomalous situation in which the overall F statistic is significant, but none of the pairwise comparisons is significant. This situation can arise because the overall F test is equivalent to a simultaneous test of the hypothesis that all possible contrasts among means are equal to zero. The contrast that is significant, however, may involve some linear combination of means such as $\mu_1 - (\mu_2 + \mu_3)/2$ rather than $\mu_1 - \mu_2$.

TUKEY'S *HSD* TEST

One of the more widely used a posteriori procedures for evaluating all pairwise comparisons among means was developed by Tukey (1953). This test, which is called the *HSD* (*honestly significant difference*) test or *WSD* (*wholly significant difference*) test, has been the subject of numerous investigations.* The test sets the experimentwise error rate at α for the collection of all pairwise comparisons. The basic assumptions of normality, homogeneity of variance, and so on, discussed at the beginning of Section 3.5 are also required. In addition, the n's in each treatment level must be equal. Alternative procedures for the unequal n and unequal variance cases are described later.

Tukey's *HSD* test is based on the sampling distribution of the studentized range statistic which, like the t distribution, was derived by William Sealey Gossett. The studentized range statistic q is

$$q = \frac{\overline{Y}_{\text{largest}} - \overline{Y}_{\text{smallest}}}{\sqrt{\frac{MS_{\text{error}}}{n}}}$$

where $\overline{Y}_{\text{largest}}$ and $\overline{Y}_{\text{smallest}}$ are the largest and smallest of p sample means, MS_{error} is an estimator of the unknown common population error variance, and $\sqrt{MS_{\text{error}}/n}$ is the standard error of the range of means, denoted by $\sigma_{\overline{Y}}$. The sampling distribution of q depends on the number of sample means used in computing the range, $\overline{Y}_{\text{largest}} - \overline{Y}_{\text{smallest}}$. It is reasonable to expect that on the average the size of the range for, say, three independent samples is larger than that for two samples and increases as the number of samples increases. This increase in the size of the range, $\overline{Y}_{\text{largest}} - \overline{Y}_{\text{smallest}}$, as the number of means increases is reflected in the studentized range distribution. Selected percentage points for the distribution of q are given in Appendix Table E.7. To enter the table, two values are required: the degrees of freedom for MS_{error} and p, the number of means on which the range, $\overline{Y}_{\text{largest}} - \overline{Y}_{\text{smallest}}$, is based. For a completely randomized design, an estimator of the population error variance is *MSWG* with $p(n - 1)$ degrees of freedom.

The critical difference, $\hat{\psi}(HSD)$, that a pairwise comparison must exceed to be declared significant is, according to Tukey's procedure,

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

where $q_{\alpha;p,\nu}$ is obtained from Appendix Table E.7 for α level of significance, p means, and ν degrees of freedom associated with MS_{error}. We will illustrate the procedure for the data in Table 3.5-1; assume that MS_{error} is equal to 28.8 with $p(n - 1) = 5(10 - 1) = 45$ degrees of freedom. The .01 level of significance will be used. For these data, $\hat{\psi}(HSD)$ corresponding to the .01 level of significance for a two-tailed test is equal to

* For a summary of research since 1953 and bibliographies see Keselman and Rogan (1977) and Miller (1977).

TABLE 3.5-1 Absolute Value of Differences Among Means ($MS_{error} = 28.8$, $p = 5$, and $n = 10$)

	$\overline{Y}_1$	$\overline{Y}_5$	$\overline{Y}_3$	$\overline{Y}_4$	$\overline{Y}_2$
$\overline{Y}_1 = 36.7$	—	3.6	6.7	10.5*	12.0*
$\overline{Y}_5 = 40.3$		—	3.1	6.9	8.4*
$\overline{Y}_3 = 43.4$			—	3.8	5.3
$\overline{Y}_4 = 47.2$				—	1.5
$\overline{Y}_2 = 48.7$					—

*$p < .01$

$$\hat{\psi}(HSD) = q_{.01;5,45}\sqrt{\frac{MS_{error}}{n}} = 4.893\sqrt{\frac{28.8}{10}} = 8.30.$$

A test of the overall null hypothesis that $\mu_1 = \mu_2 = \cdots = \mu_p$ is provided by a comparison of the largest pairwise difference between means, $\hat{\psi} = \overline{Y}_{largest} - \overline{Y}_{smallest}$, with the critical difference $\hat{\psi}(HSD)$. This test procedure, which utilizes a range statistic, is an alternative to the overall F test. For most sets of data, the range and F tests lead to the same decision concerning the overall null hypothesis. However, the F test is generally more powerful. According to Table 3.5-1, the difference between the largest and smallest means is equal to 12.0. Because this difference exceeds $\hat{\psi}(HSD) = 8.30$, the overall null hypothesis is rejected. An examination of Table 3.5-1 indicates that three pairwise comparisons—those starred—exceed the critical difference and hence are declared significant at the .01 level. It should be noted that the values of $q_{\alpha;p,\nu}$ in Appendix Table E.7 are appropriate for testing two-tailed hypotheses; this is true for all of the a posteriori procedures described in Section 3.5.

It is instructive to compare the critical difference for Tukey's *HSD* test with those for the *LSD*, Dunn, and Dunn-Šidák tests. The critical differences are

$$\hat{\psi}(HSD) = q_{.01;5,45}\sqrt{\frac{MS_{error}}{n}} = 4.893(1.697) = 8.30$$

$$\hat{\psi}(LSD) = t_{.01/2,45}\sqrt{\frac{2MS_{error}}{n}} = 2.689(2.400) = 6.45$$

$$\hat{\psi}(D) = tD_{.01/2;10,45}\sqrt{\frac{2MS_{error}}{n}} = 3.520(2.400) = 8.45$$

$$\hat{\psi}(DS) = tDS_{.01/2;10,45}\sqrt{\frac{2MS_{error}}{n}} = 3.519(2.400) = 8.45.$$

As expected, the *LSD* test is the most powerful because at the second stage of the testing procedure it does not control the error rate at α for the collection of tests. The least sensitive procedures in this example are those of Dunn and Dunn-Šidák. They become more powerful relative to Tukey's *HSD* test as the number of comparisons among the p means is reduced. For example, if an experimenter had planned to make only eight instead of all ten pairwise comparisons among means, the critical difference for the Dunn-Šidák procedure would have been only

$$\hat{\psi}(DS) = tDS_{.01/2;8,45}\sqrt{\frac{2MS_{\text{error}}}{n}} = 3.443(2.400) = 8.26$$

which is less than that for Tukey's procedure.

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

$$\hat{\psi}_i - \hat{\psi}(HSD) \leq \psi_i \leq \hat{\psi}_i + \hat{\psi}(HSD)$$

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

$$\hat{\psi}_i = \overline{Y}_{.j} - \overline{Y}_{.j'}$$

$$\psi_i = \mu_j - \mu_{j'}.$$

The test statistic for Tukey's procedure is

$$q = \frac{\hat{\psi}_i}{\hat{\sigma}_{\overline{Y}}} = \frac{\overline{Y}_{.j} - \overline{Y}_{.j'}}{\sqrt{\frac{MS_{\text{error}}}{n}}}.$$

If $|q| \geq q_{\alpha;p,\nu}$, the null hypothesis, H_0: $\mu_j - \mu_{j'} = 0$, is rejected. Tukey's procedure can also be used with the conventional t statistic (3.2-1) described in Section 3.2.

$$t = \frac{\hat{\psi}_i}{\hat{\sigma}_{\psi_i}} = \frac{\overline{Y}_{.j} - \overline{Y}_{.j'}}{\sqrt{\frac{2MS_{\text{error}}}{n}}}$$

The critical value for this t statistic is $q_{\alpha;p,\nu}/\sqrt{2}$, since $t = q/\sqrt{2}$.

Tukey's procedure can be extended to test nonpairwise contrasts. However, this is not recommended since for this application it is less powerful than Scheffé's procedure, which is described later.

Earlier we noted that Tukey's procedure assumes equal n's. This is necessary in order for the sample means to have the same variance σ^2/n. The procedure also assumes that the p population variances are homogeneous. In the following sections we will describe procedures for evaluating pairwise comparisons that do not require these assumptions.

SPJØTVOLL AND STOLINE'S MODIFICATION OF THE HSD TEST

Over the years a variety of a posteriori procedures have been suggested for evaluating pairwise comparisons when the sample n's are unequal: Dunn (1974), Gabriel (1978b), Hochberg (1974), Kramer (1956), Šidák (1967), Scheffé (1953), Spjøtvoll and Stoline (1973), and Tukey (1953). Ury (1976) compared the procedures of Dunn-Šidák (Šidák, 1967), Hochberg (1974), Scheffé (1953), and Spjøtvoll and Stoline (1973). He concluded that when the n's are similar the preferred procedure was that due to Spjøtvoll and Stoline. When the n's are quite different or a very high

level of significance is adopted ($\alpha < .01$), one of the other procedures may be preferred (Ury, 1976; Stoline, 1978). Probably the most common situation involving unequal n's in the behavioral sciences and education is one in which the n's are similar.

The Spjøtvoll-Stoline test, which is a generalization of Tukey's test, is appropriate for unequal n's and is referred to as the T' test. The basic assumptions of normality, homogeneity of variance, and so on, discussed in Section 3.5 are required for this test. Of course the test should be preceded by a significant test of the overall null hypothesis. The T' test is based on the studentized augmented range distribution.* The critical difference $\hat{\psi}(T')$ that a pairwise comparison must exceed to be declared significant is

$$\hat{\psi}(T') = q'_{\alpha;p,\nu}\sqrt{\frac{MS_{\text{error}}}{n_{\min}}}$$

where $q'_{\alpha;p,\nu}$ is obtained from Appendix Table E.18 for α level of significance, p means, and ν degrees of freedom associated with MS_{error}; and $n_{\min}$ is the minimum of n_j and $n_{j'}$, the sample sizes used to compute $\overline{Y}_j$ and $\overline{Y}_{j'}$, respectively. When p exceeds eight treatment means, the value of $q'_{\alpha;p,\nu}$ can be obtained from the studentized range distribution (Appendix Table E.7).

Simultaneous $100(1 - \alpha)\%$ confidence intervals for the Spjøtvoll-Stoline procedure are given by

$$\hat{\psi}_i - \hat{\psi}(T') \leq \psi_i \leq \hat{\psi}_i + \hat{\psi}(T')$$

where $\hat{\psi}(T') = q'_{\alpha;p,\nu}\sqrt{\dfrac{MS_{\text{error}}}{n_{\min}}}$

$$\hat{\psi}_i = \overline{Y}_j - \overline{Y}_{j'}$$

$$\psi_i = \mu_j - \mu_{j'}.$$

The test statistic for the Spjøtvoll-Stoline procedure is

$$q'T' = \frac{\hat{\psi}_i}{\hat{\sigma}_{\overline{Y}}} = \frac{\overline{Y}_j - \overline{Y}_{j'}}{\sqrt{\dfrac{MS_{\text{error}}}{n_{\min}}}}.$$

If $|q'T'| \geq q'_{\alpha;p,\nu}$, the null hypothesis, H_0: $\mu_j - \mu_{j'} = 0$, is rejected.

TUKEY-KRAMER MODIFICATION OF THE *HSD* TEST

Tukey (1953) and Kramer (1956) independently proposed a modification of the *HSD* test for the case in which the sample n's are unequal but the basic assumptions of normality, homogeneity of variance, and so on are tenable. We will refer to the test as the Tukey-Kramer procedure. For this test, the critical difference, $\hat{\psi}(TK)$, that a pairwise comparison must exceed to be declared significant is

* See Scheffé (1959, 78) for a description of the studentized augmented range.

$$\hat{\psi}(TK) = q_{\alpha;p,\nu}\sqrt{MS_{\text{error}}\left(\frac{1}{n_j} + \frac{1}{n_{j'}}\right)\Big/ 2}$$

where $q_{\alpha;p,\nu}$ is obtained from the studentized range distribution (Appendix Table E.7) for α level of significance, p means, and ν degrees of freedom associated with MS_{error}.

The Tukey-Kramer procedure is thought by some to be liberal, that is, to not control the experimentwise error rate at or less than α. Research by Dunnett (1980) and unpublished research cited by Dunnett (1980), however, suggests that the procedure is conservative at least for experiments involving moderate to severe imbalance among sample n's. Furthermore the size of the critical difference, $\hat{\psi}(TK)$, is always less than that for the Spjøtvoll-Stoline procedure. Pending further research, the Spjøtvoll-Stoline procedure is recommended when sample n's are approximately equal; the Tukey-Kramer procedure is recommended when there is a moderate or large imbalance among the sample n's. We turn next to two procedures that are appropriate for experiments with unequal population variances and equal or unequal sample sizes.

ROBUST PROCEDURES FOR A POSTERIORI NONORTHOGONAL PAIRWISE COMPARISONS

The problem of evaluating pairwise comparisons when the variances are unequal or both the variances and the n's are unequal has also received considerable attention in recent years.* The most promising a posteriori procedures are those of Games and Howell (1976) and Tamhane (1977, 1979). Both procedures use the Behrens-Fisher statistic described in Section 3.2 to estimate the standard error of a contrast, σ_{ψ_i}, and the Welch procedure for determining the degrees of freedom for the standard error of the contrast. However, the Games-Howell procedure uses the studentized range distribution while the Tamhane procedure uses Student's t distribution and the Šidák multiplicative inequality. Since the procedures are used for data snooping, they should be preceded by a significant test of the overall null hypothesis.

The critical difference $\hat{\psi}(GH)$ that a pairwise comparison must exceed in order to reject the hypothesis H_0: $\mu_j - \mu_{j'} = 0$ for the Games and Howell procedure is

$$\hat{\psi}(GH) = q_{\alpha;p,\nu'}\sqrt{\left(\frac{\hat{\sigma}_j^2}{n_j} + \frac{\hat{\sigma}_{j'}^2}{n_{j'}}\right)\Big/ 2}$$

where $q_{\alpha;p,\nu'}$ is obtained from the studentized range distribution in Appendix Table E.7 for α level of significance, p means, and degrees of freedom equal to

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

* See, for example, Brown and Forsythe (1974); Dalal (1978); Games and Howell (1976); Hochberg (1976); Keselman and Rogan (1977); Keselman, Games, and Rogan (1979); Miller (1966, 43); Spjøtvoll (1972); Tamhane (1977, 1979); and Ury and Wiggins (1971).

The critical difference, $\hat{\psi}(T2)$, for the Tamhane procedure is

$$\hat{\psi}(T2) = tDS_{\alpha/2;C,\nu'} \sqrt{\frac{\hat{\sigma}_j^2}{n_j} + \frac{\hat{\sigma}_{j'}^2}{n_{j'}}}$$

where $tDS_{\alpha/2;C,\nu'}$ is obtained from the t distribution using the Šidák multiplicative inequality (Appendix Table E.17) for $C = p(p-1)/2$ pairwise comparisons and degrees of freedom, ν', as defined for the Games-Howell procedure. If at least one of the following conditions holds, ν' is set equal to $n_j + n_{j'} - 2$ (Tamhane, 1979; Ury and Wiggins, 1971).

1. $9/10 \leq n_j/n_{j'} \leq 10/9$
2. $9/10 \leq (\hat{\sigma}_j^2/n_j)/(\hat{\sigma}_{j'}^2/n_{j'}) \leq 10/9$
3. $4/5 \leq n_j/n_{j'} \leq 5/4$ and $1/2 \leq (\hat{\sigma}_j^2/n_j)/(\hat{\sigma}_{j'}^2/n_{j'}) \leq 2$
4. $2/3 \leq n_j/n_{j'} \leq 3/2$ and $3/4 \leq (\hat{\sigma}_j^2/n_j)/(\hat{\sigma}_{j'}^2/n_{j'}) \leq 4/3$

The procedures of Tamhane and Games and Howell appear to be excellent choices for evaluating pairwise comparisons when the variances or the variances and sample n's are unequal. Monte Carlo sampling studies indicate that even when the unknown variances are homogeneous, use of the procedures does not lead to a substantial loss of power (Keselman, Games, and Rogan, 1979). The Games-Howell procedure is slightly more powerful than the Tamhane procedure (Tamhane, 1979). However, the Tamhane procedure controls the experimentwise error rate at α or less while the Games-Howell procedure can become slightly liberal (Keselman, Games, and Rogan, 1979; Tamhane, 1979). A choice between them depends on whether one prefers slightly greater power or better control of the experimentwise error rate.

SCHEFFÉ'S *S* TEST

The Scheffé S procedure (1953) is one of the most flexible, conservative, and robust data snooping procedures available. If the overall F statistic is significant, Scheffé's procedure can be used to evaluate all a posteriori contrasts among means, not just the pairwise comparisons. In addition, it can be used with unequal n's. The error rate experimentwise is equal to α for the infinite number of possible contrasts among $p \geq 3$ means. Since an experimenter always evaluates a subset of the possible contrasts, Scheffé's procedure tends to be conservative. It is much less powerful than Tukey's *HSD* procedure for evaluating pairwise comparisons, for example, and consequently is recommended only when complex contrasts are of interest. Scheffé's procedure uses the F sampling distribution and, like ANOVA, is robust with respect to nonnormality and heterogeneity of variance.

The critical difference $\hat{\psi}(S)$ that a contrast must exceed to be declared significant is given by

$$\hat{\psi}(S) = \sqrt{(p-1)F_{\alpha;\nu_1,\nu_2}} \sqrt{MS_{\text{error}} \sum_{j=1}^{p} \frac{c_j^2}{n_j}}$$

where p is the number of means in the experiment (family) and $F_{\alpha;\nu_1,\nu_2}$ is obtained from Appendix Table E.5 for ν_1 equal to $p-1$ degrees of freedom and ν_2 equal to the

degrees of freedom associated with MS_{error}. For purposes of comparison, the value of $\hat{\psi}(S)$ will be computed for a pairwise comparison using the data in Table 3.5-1. We will assume that the overall null hypothesis $\mu_1 = \mu_2 = \cdots = \mu_p$ has been rejected using an F statistic. The value of the critical difference is

$$\hat{\psi}(S) = \sqrt{(5-1)3.77}\sqrt{28.8\left[\frac{(1)^2}{10} + \frac{(-1)^2}{10}\right]} = 3.883(2.400) = 9.32.$$

The critical difference for Tukey's procedure is only $\hat{\psi}(HSD) = 8.30$. Hence, if one wanted to evaluate only pairwise comparisons, Scheffé's procedure would be a poor choice. The advantage of Scheffé's procedure, the ability to evaluate all possible contrasts, comes at a price—low power.

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

$$\hat{\psi}_i - \hat{\psi}(S) \leq \psi_i \leq \hat{\psi}_i + \hat{\psi}(S)$$

where $\hat{\psi}(S) = \sqrt{(p-1)F_{\alpha;\nu_1,\nu_2}}\sqrt{MS_{\text{error}}\sum_{j=1}^{p}\frac{c_j^2}{n_j}}$

$$\hat{\psi}_i = \sum_{j=1}^{p} c_j \overline{Y}_j$$

$$\psi_i = \sum_{j=1}^{p} c_j \mu_j.$$

The test statistic for Scheffé's procedure is

$$F = \frac{\hat{\psi}_i^2}{\hat{\sigma}_{\hat{\psi}_i}^2} = \frac{\left(\sum_{j=1}^{p} c_j \overline{Y}_j\right)^2}{MS_{\text{error}}\sum_{j=1}^{p}\frac{c_j^2}{n_j}}.$$

To be significant, F must exceed F', where $F' = (p-1)F_{\alpha;\nu_1,\nu_2}$.

BROWN-FORSYTHE *BF* PROCEDURE

A procedure, which is a modification of Scheffé's S test, for evaluating all a posteriori contrasts among means following a significant F test of the overall null hypothesis has been described by Brown and Forsythe (1974). Their procedure is even more robust with respect to heterogeneity of variance than Scheffé's. This is accomplished by using the Behrens-Fisher statistic described in Section 3.2 to estimate the standard error of a contrast σ_{ψ_i} and the Welch (1947) procedure for determining the degrees of freedom for the standard error of the contrast. Like Scheffé's procedure, it can be used with unequal sample n's. The experimentwise error rate is less than or equal to α.

The critical difference $\hat{\psi}(BF)$ that a contrast must exceed in order to reject the null hypothesis for the population contrast is given by

$$\hat{\psi}(BF) = \sqrt{(p-1)F_{\alpha;\nu_1,\nu_2'}} \sqrt{\frac{c_1^2\hat{\sigma}_1^2}{n_1} + \frac{c_2^2\hat{\sigma}_2^2}{n_2} + \cdots + \frac{c_p^2\hat{\sigma}_p^2}{n_p}}$$

where $F_{\alpha;\nu_1,\nu_2'}$ is obtained from Appendix Table E.5 for ν_1 equal to $p-1$ degrees of freedom and ν_2' equal to

$$\nu_2' = \frac{\left(\sum_{j=1}^{p} \frac{c_j^2\,\hat{\sigma}_j^2}{n_j}\right)^2}{\sum_{j=1}^{p} \frac{c_j^4\,\hat{\sigma}_j^4}{n_j^2\,(n_j-1)}}$$

degrees of freedom.

If an experimenter is interested in evaluating only pairwise comparisons and the population variances are believed to be heterogeneous, the procedures of Games and Howell or Tamhane described earlier should be used since for this case they are more powerful.

NEWMAN-KEULS TEST

A different approach to evaluating a posteriori pairwise comparisons stems from the work of Student (1927), Newman (1939), and Keuls (1952). The Newman-Keuls procedure is based on a stepwise or layer approach to significance testing. Sample means are ordered from the smallest to the largest, as in Table 3.5-2. The largest difference, which involves means that are $r = p$ steps apart, is tested first at α level of significance; if significant, means that are $r = p - 1$ steps apart are tested at α level of significance and so on. The Newman-Keuls procedure provides an r-mean significance level equal to α for each group of r ordered means; that is, the probability of falsely rejecting the hypothesis that all means in an ordered group are equal is α. It follows that the concept of error rate applies neither on an experimentwise nor on a per comparison basis—the actual error rate falls somewhere between the two. The

TABLE 3.5-2 Absolute Value of Differences Among Means ($MS_{error} = 28.8$, $p = 5$, and $n = 10$)

	$\bar{Y}_{\cdot 1}$	$\bar{Y}_{\cdot 5}$	$\bar{Y}_{\cdot 3}$	$\bar{Y}_{\cdot 4}$	$\bar{Y}_{\cdot 2}$
$\bar{Y}_{\cdot 1} = 36.7$	–	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
$\bar{Y}_{\cdot 2} = 48.7$					–

*$p < .01$

Newman-Keuls procedure, like Tukey's procedure, requires equal sample n's. It differs from all of the tests described previously in that it cannot be used to construct confidence intervals. The stepwise or multistage character of the test has no counterpart in confidence intervals.

The critical difference $\hat{\psi}(W_r)$ that two means separated by r steps must exceed to be declared significant is, according to the Newman-Keuls procedure,

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

where $q_{\alpha;r,\nu}$ is obtained from Appendix Table E.7 for α level of significance, means separated by r steps, and ν degrees of freedom associated with MS_{error}. The subscript r designates the number of steps separating ordered means. Consider the following means from Table 3.5-2.

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

Mean one is defined as being five steps away from mean two, four steps away from mean four, and so on. Two values are required to enter the studentized range table—the degrees of freedom for MS_{error} and r. For $p(n-1) = 45$ degrees of freedom, the respective values of $q_{.01;r,45}$ from Appendix Table E.7 are

$$q_{.01;2,45} = 3.800, \quad q_{.01;3,45} = 4.340, \quad q_{.01;4,45} = 4.663, \quad q_{.01;5,45} = 4.893.$$

The critical differences, $\hat{\psi}(W_r)$, are the products of

$$q_{.01;r,45}\sqrt{\frac{MS_{\text{error}}}{n}}$$

where $MS_{\text{error}} = 28.8$
$n = 10.$

$$\hat{\psi}(W_2) = 3.800\sqrt{\frac{28.8}{10}} = 6.45$$

$$\hat{\psi}(W_3) = 4.340\sqrt{\frac{28.8}{10}} = 7.37$$

$$\hat{\psi}(W_4) = 4.663\sqrt{\frac{28.8}{10}} = 7.91$$

$$\hat{\psi}(W_5) = 4.893\sqrt{\frac{28.8}{10}} = 8.30$$

The critical difference is the difference that two means r steps apart must exceed in order to reject the null hypothesis H_0: $\mu_j - \mu_{j'} = 0$.

There is a prescribed sequence in which tests on pairwise comparisons must be performed. As noted earlier, the means must first be arranged in order of increasing size as in Table 3.5-2. The first test is made in row one for $\overline{Y}_1$ versus $\overline{Y}_2$. Because these two means are separated by five steps, the critical difference that this comparison must exceed is 8.30. The difference in Table 3.5-2 exceeds this critical difference. This first test is a range test of the overall null hypothesis. If it had been insignificant, no other comparisons would have been tested. The next comparison tested is $\overline{Y}_1$ versus $\overline{Y}_4$ in row one. The critical difference for this comparison is 7.91. This comparison is also significant. This procedure is continued in row one until a nonsignificant comparison is found. In this example, the comparison $\overline{Y}_1$ versus $\overline{Y}_3$ does not exceed 7.37, so no further tests are made in this row or in the following rows to the left of the column headed by $\overline{Y}_4$. The next test is the comparison of $\overline{Y}_5$ versus $\overline{Y}_2$ in row two. The critical difference for this test is 7.91. This process is repeated for each successive row until a nonsignificant comparison is found or until the column at which tests were stopped in the preceding row is reached. In this example, the Newman-Keuls procedure leads to three significant pairwise comparisons—those starred in Table 3.5-2. This is the same number that was declared significant by Tukey's *HSD* procedure but one more than was obtained by Scheffé's procedure.

It should be noted that the Newman-Keuls and Tukey procedures require the same critical difference, 8.30, for the first comparison that is tested. The Tukey procedure uses this critical difference for all of the remaining tests while the Newman-Keuls procedure reduces the size of the critical difference, depending on the number of steps separating the ordered means. As a result, the Newman-Keuls test is more powerful than Tukey's test. Remember, however, that the Newman-Keuls procedure does not control the experimentwise error rate at α.

Frequently a test of the overall null hypothesis $\mu_1 = \mu_2 = \cdots = \mu_p$ is performed with an F statistic in ANOVA rather than with a range statistic. If the F statistic is significant, Shaffer (1979) recommends using the critical difference $\hat{\psi}(W_{r-1})$ instead of $\hat{\psi}(W_r)$ to evaluate the largest pairwise comparison at the first step of the testing procedure. The testing procedure for all subsequent steps is unchanged. She has shown that the modified procedure leads to greater power at the first step without affecting control of the type I error rate. This makes *dissonances*, in which the overall null hypothesis is rejected by an F test without rejecting any one of the proper subsets of comparisons, less likely.

DUNCAN'S NEW MULTIPLE RANGE TEST

Duncan (1955) has developed a procedure for evaluating a posteriori pairwise comparisons that shares many of the features of the Newman-Keuls procedure including a stepwise approach to significance testing, prescribed sequence for performing tests, and absence of confidence interval procedures. The r-mean significance level for this test is equal to $1 - (1 - \alpha)^{r-1}$. For five ordered means in a group and $\alpha = .01$, the probability of falsely rejecting the hypothesis that all means in the ordered group are equal is $1 - (1 - .01)^{5-1} = .0394$. The corresponding r-mean significance level for

the Newman-Keuls procedure is .01 and remains at .01 for each group of ordered means. For Duncan's procedure, the r-mean significance level decreases as r increases, for example

$$\begin{aligned} &\text{2-means} && 1 - (1 - .01)^{2-1} = .01 \\ &\text{3-means} && 1 - (1 - .01)^{3-1} = .0199 \\ &\text{4-means} && 1 - (1 - .01)^{4-1} = .0297 \\ &\text{5-means} && 1 - (1 - .01)^{5-1} = .0394. \end{aligned}$$

Duncan (1955) has argued that if $p > 2$, an experiment is more likely to contain significant contrasts than if the experiment contains only two means. Thus, as p increases, the test should become more powerful. He reasons further that p means can be used to form $p - 1$ orthogonal contrasts, each of which can be tested at α level of significance. We saw in Section 3.3 that for $p - 1$ orthogonal contrasts, the error rate experimentwise is $1 - (1 - \alpha)^{p-1}$. Duncan's procedure provides the same r-mean significance level as that tolerated for $r - 1$ orthogonal contrasts.

The critical difference $\hat{\psi}(W_r)$ that two means separated by r steps must exceed to be declared significant is, according to Duncan's procedure,

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

where $q_{\alpha;r,\nu}$ is obtained from Duncan's table in the Appendix (Table E.8) for α level of significance, means separated by r steps, and ν degrees of freedom associated with MS_{error}. For the data in Table 3.5-2, where $MS_{\text{error}} = 28.8$ and $n = 10$, the critical differences are

$$\hat{\psi}(W_2) = q_{.01;2,45} \sqrt{\frac{28.8}{10}} = 3.800(1.697) = 6.45$$

$$\hat{\psi}(W_3) = q_{.01;3,45} \sqrt{\frac{28.8}{10}} = 3.967(1.697) = 6.73$$

$$\hat{\psi}(W_4) = q_{.01;4,45} \sqrt{\frac{28.8}{10}} = 4.077(1.697) = 6.92$$

$$\hat{\psi}(W_5) = q_{.01;5,45} \sqrt{\frac{28.8}{10}} = 4.153(1.697) = 7.05.$$

The sequence in which the tests are carried out is identical to that for the Newman-Keuls procedure. An examination of Table 3.5-2 reveals that the null hypotheses for four of the pairwise comparisons can be rejected: $\mu_1 - \mu_2$, $\mu_1 - \mu_4$, $\mu_5 - \mu_2$, and $\mu_5 - \mu_4$. This is one more significant test than was obtained using the Newman-Keuls and Tukey *HSD* procedures.

An extension of Duncan's new multiple range test for the case of unequal sample n's is described by Kramer (1956). Other applications of the test are discussed by Duncan (1957). Scheffé (1959, 718) has criticized the justification originally advanced for Duncan's test.

3.6 COMPARISON OF MULTIPLE COMPARISON PROCEDURES

A variety of multiple comparison procedures has been described in this chapter. The procedures and their salient characteristics are summarized in Table 3.6-1. Most of the procedures have been illustrated using the same set of data involving five means. An indication of the relative power of the procedures can be obtained by comparing the size of their critical difference. For purposes of this comparison, we will assume that the multiple t procedure and Dunnett's procedure are used to evaluate $p - 1$ a priori contrasts and the other procedures are used to evaluate all ten pairwise comparisons among the five means. This comparison is not fair to the Dunn and Dunn-Šidák procedures because they were designed for evaluating C a priori contrasts, nor is it fair to the Scheffé procedure which is not recommended for evaluating pairwise contrasts. Nevertheless an examination of the critical differences in Table 3.6-2 is instructive.

The relative merits of various multiple comparison procedures has engendered much debate among statisticians in recent years. Each of the procedures in Table 3.6-1 has been recommended by one or more statisticians (Carmer and Swanson, 1973; Dunnett, 1980; Einot and Gabriel, 1975; Gabriel, 1978a; Games and Howell, 1976; Keselman, Games, and Rogan, 1979; Keselman and Rogan, 1977, 1978; Ramsey, 1978; Tamhane, 1979; Thomas, 1974; Ury, 1976). Several statisticians (Games, 1971; Hopkins and Anderson, 1973; Hopkins and Chadbourn, 1967) have developed helpful guides in the form of flowcharts for selecting a multiple comparison procedure. Here too one finds little agreement. The problem facing an experimenter 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 an experimenter can make informed choices.

3.7 REVIEW EXERCISES

1. [3.1] 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?

†2. [3.1] Distinguish between a pairwise comparison and a contrast.

3. [3.1] 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) μ_1 versus mean of μ_2, μ_3, and μ_4 **†d)** Mean of μ_1 and μ_2 versus mean of μ_3 and μ_4

e) Mean of μ_1 and μ_2 versus mean of μ_3, μ_4, and μ_5

†f) μ_1 versus the weighted mean of μ_2 and μ_3, where μ_2 is weighted twice as much as μ_3

TABLE 3.6-1 Summary of Characteristics of Multiple Comparison Procedures

	Test is appropriate for							
	Pairwise Comparisons Only	*$p - 1$ Pairwise Comparisons*	*Complex Comparisons*	*C Complex Comparisons*	*Equal n's Only*	*Un-equal n's*	*Homo-geneous Variances*	*Hetero-geneous Variances*
	A Priori Orthogonal Contrasts							
Multiple *t*			x			x	x	
Multiple *F*			x			x	x	
Multiple *t* with Behrens-Fisher and Welch Procedures			x			x		x
	A Priori Nonorthogonal Contrasts							
Dunn *D*				x		x	x	
Dunn-Šidák *DS*				x		x	x	
Dunnett *D'*		x				x	x	x*
	A Posteriori Nonorthogonal Contrasts							
Fisher *LSD*	x					x	x	
Tukey *HSD*	x				x		x	
Spjøtvoll-Stoline *T'*	x					x	x	
Tukey-Kramer *TK*	x					x	x	
Games-Howell *GH*	x					x		x
Tamhane *T2*	x					x		x
Scheffé *S*			x			x	x	
Brown-Forsythe *BF*			x			x		x
Newman-Keuls	x				x		x	
Duncan	x				x		x	

*With modification

TABLE 3.6-2 Comparison of Critical Difference for Several Multiple Comparison Procedures

Test	*Critical Difference*				
	A Priori Orthogonal Contrasts				
Multiple t	6.45 for $p - 1 = 4$ contrasts				
	A Priori Nonorthogonal Contrasts				
Dunnett	7.60 for $p - 1 = 4$ contrasts with a control group mean				
Dunn-Šidák	8.45 for $C = 10$ pairwise comparisons				
Dunn	8.45 for $C = 10$ pairwise comparisons				
	A Posteriori Nonorthogonal Contrasts				
	Number of steps separating means				
	2	3	4	5	
Fisher *LSD*	6.45	same	same	same	for 10 pairwise comparisons
Duncan	6.45	6.73	6.92	7.05	
Newman-Keuls	6.45	7.37	7.91	8.30	
Tukey *HSD*	8.30	same	same	same	
Scheffé *S*	9.32	same	same	same	

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. [3.1] Which of the following meet the requirements for a contrast?

†**a)** $\mu_1 - \mu_2$ †**b)** $2\mu_1 - \mu_2 - \mu_3$ †**c)** $\mu_1 - (1/3)\mu_2 - (1/3)\mu_3$

d) $(1\,1/2)\mu_1 - \mu_2 - (1\,1/2)\mu_3$ **e)** $\mu_1 - (1/4)\mu_2 - (1/4)\mu_3 - (1/4)\mu_4$

†**f)** $(3/4)\mu_1 - (1/4)\mu_2 - (1/4)\mu_3 - (1/4)\mu_4$

g) $(1/2)\mu_1 + (1/2)\mu_2 - (1/3)\mu_3 - (1/3)\mu_4 - (1/3)\mu_5$

5. [3.1] Which contrasts in Exercise 4 satisfy $|c_1| + |c_2| + \cdots + |c_p| = 2$?

6. [3.1] Indicate the number of pairwise comparisons that can be constructed for the following designs.

†**a)** Type CR-3 design **b)** Type CR-4 design

†**c)** Type CR-5 design **d)** Type CR-6 design

7. [3.1] Which of the following sets of contrasts are orthogonal? Assume that the n's are equal.

c) Compute the correlations among the contrasts; assume that $c_{1j} = 1\ -1\ 0\ 0$, $c_{2j} = 1\ 0\ -1\ 0$, et cetera.

d) Compare the relative efficiency of Dunn's procedure with that of the Dunn-Šidák procedure.

e) Suppose that the experimenter is only interested in the $p - 1 = 3$ contrasts involving treatment level a_4. For this case, compare the relative efficiency of the Dunn and Dunn-Šidák procedures with Dunnett's procedure.

†19. [3.5] Exercise 17 described an experiment to evaluate the effects of four dosages of ethylene glycol on the reaction time of chimpanzees.

a) Use Tukey's procedure to test the overall 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 3.5-1. Let $\alpha_{EW} = .05$.

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

c) If you did Exercise 17(d), compare the relative efficiency of the Tukey and Dunn-Šidák procedures.

d) Use the Newman-Keuls test to evaluate all pairwise comparisons.

e) Use Duncan's new multiple range test to evaluate all pairwise comparisons.

20. [3.5] 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 overall 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 3.5-1. Let $\alpha_{EW} = .05$.

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

c) If you did Exercise 18(d), compare the relative efficiency of the Tukey and Dunn-Šidák procedures.

d) Use the Newman-Keuls test to evaluate all pairwise comparisons.

e) Use Duncan's new multiple range test to evaluate all pairwise comparisons.

†21. [3.5] Exercise 10 described an experiment to evaluate the effectiveness of three approaches to drug education in junior high school. Assume that the overall null hypothesis was rejected at the .05 level of significance.

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

$$\mu_1 - \mu_3 = 0, \qquad \mu_2 - \mu_3 = 0, \qquad \text{and} \qquad (\mu_1 + \mu_2)/2 - \mu_3 = 0.$$

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

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 given null hypotheses.

22. [3.5] The effects of simulator training involving synergistic 6-degree-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 ten sorties with platform motion in the Advanced Simulator for Pilot Training; those in group a_2 also received ten 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 overall null hypothesis was rejected at the .05 level.

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

$$\mu_1 - \mu_2 = 0 \qquad \text{and} \qquad (\mu_1 + \mu_2)/2 - \mu_3 = 0.$$

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

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 given null hypotheses.