Multiple Comparisons for Orthogonal Contrasts: Examples and Tables

Robert E. Bechhofer
School of Operations Research
and Industrial Engineering
Cornell University
Ithaca, NY 14853

Charles W. Dunnett
Dept. of Clinical Epidemiology
and Biostatistics, and
Dept. of Mathematical Sciences
McMaster University
Hamilton, Ontario, CANADA L8S 3Z5

In many experimental situations the pertinent inferences are made on the basis of orthogonal contrasts among the treatment means (as in 2^k factorial experiments). In this setting a particularly useful form of inference is one involving multiple comparisons. The present article describes situations in which such inferences are meaningful, gives examples of their use, and provides a table of constants needed to implement such multiple comparison procedures. The procedures can also be used for statistically legitimate “data snooping” (in the sense of Scheffe 1959, p. 80) to help decide which contrasts within a specified set warrant further study.

KEY WORDS: Multiple comparisons; Orthogonal contrasts; Joint confidence intervals; Experimentwise error rates; Studentized maximum modulus; Simultaneous inference.

1. INTRODUCTION

Most experiments are planned in such a way that several questions can be asked of the data. For instance, in a k-treatment experiment there may be several comparisons between individual treatment means and between combinations of treatment means that are of interest to the experimenter. Thus, multiple significance tests may be performed or multiple confidence interval estimates determined from the same set of experimental data. Such practices should not be discouraged, of course, since it is understandable that an experimenter would wish to extract as much information as possible from a given set of data. On the other hand, if a multiplicity of tests is applied to the same set of data, the chances of finding one or more interesting or unusual effects are obviously increased (even if, in fact, no true differences exist). Before deciding to take some action on the basis of such indicated findings, for example to investigate them further or to report them in a scientific journal, the experimenter may wish to make some allowance in the test procedure for the fact that multiple tests have been carried out. As an extreme example, suppose that several tests of significance are performed to examine various treatment effects of interest, and exactly one of these turns out to be statistically significant in the usual sense. How can the experimenter take into account (in terms of a P value) the fact that the multiplicity of tests carried out on the data resulted in the selection of this particular effect as being the “most statistically significant one” in the set? Alternatively, suppose that there are several treatment effects for which confidence interval estimates are to be determined. If a conclusion is to be reached on the basis of the entire set of simultaneous confidence intervals, how much do the confidence intervals need to be widened to achieve a specified joint confidence coefficient for the entire set? It is to answer such questions that multiple test procedures and multiple comparison procedures have been developed.

In this article we consider experiments in which the results can be analyzed in terms of meaningful orthogonal contrasts among the treatment means. This is the situation, for example, in 2^k factorial experiments, and in experiments with quantitative factors for which orthogonal polynomials (Fisher and Yates 1938–1963; Davies 1978, Appendix 8C) are employed to fit a regression curve or surface. Nair (1948, pp. 21 and 27) considers such analyses. Cochran and Cox (1957, Sec. 3.4–3.5) discuss orthogonal contrasts in some detail. Ackermann (1979) proved a general formula to calculate the values of orthogonal polynomials for the case of nonequidistant levels and unequal numbers of observations.

For hypothesis testing the orthogonality of the treatment contrasts makes it possible to partition the
sum of squares for treatments into a set of one-degree-of-freedom sums of squares that add up to the sum of squares for treatments; under normality each individual sum of squares associated with one of the treatment contrasts is distributed as \(\sigma^2 \chi^2\) with one degree of freedom, independently of all of the other individual sums of squares. The orthogonality of the treatment contrasts also guarantees that their usual best linear unbiased estimators, which are normally distributed, are uncorrelated.

If only one contrast is of interest, then Student's \(t\) can be used for hypothesis testing or for interval estimation. However, when two or more hypotheses are each tested separately at level of significance \(\alpha\) using (say) Student's \(t\), the experimentwise error rate (i.e., the probability of at least one false positive) is greater than \(\alpha\). To help compensate for this effect some experimenters use, for each test, a common value of \(\alpha\) that is smaller than the one that would typically be employed if only one hypothesis were being tested; others use critical values based on Bonferroni inequalities. (The problem is further complicated because the individual tests are not independent since each usually employs the same residual mean square as the estimator of the underlying variance.) Alternatively, as we advocate and illustrate here, one can use exact critical values based on a specified experimentwise error rate.

In this article we mainly consider joint two-sided confidence interval estimation of orthogonal contrasts rather than joint hypothesis testing. In any case, joint interval estimation can easily be reformulated as joint hypothesis testing. We describe situations in which the joint-interval-estimation approach would appear to be appropriate and show how to make exact confidence statements concerning the joint interval estimates of the orthogonal contrasts that are of interest.

A table of constants needed to implement the procedure has been prepared. This table is an abridged version of more extensive tables given in Bechhofer and Dunnett (1981). The constants, which are based on a special case of the multivariate Student's \(t\) of Dunnett and Sobel (1954) and Cornish (1954), may also be used in situations where a set of contrasts that are not orthogonal is of interest to the experimenter. In that case, the joint confidence intervals are no longer exact, but they are conservative in the sense that their joint confidence coefficient equals or exceeds the nominal \(1 - \alpha\). Thus they may be used in place of the constants obtained by the Bonferroni method and they are superior to those obtained by the Bonferroni method since they provide narrower intervals.

Section 2 contains a brief outline of the theory underlying the method. A more complete account is given in Bechhofer and Dunnett (1981). Section 3 describes some examples of the application of the tables.

### 2. OUTLINE OF THE THEORY AND DESCRIPTION OF TABLES

#### 2.1 Basic Theory

We assume the usual one-way ANOVA model, namely, \(Y_{ij} = \mu_i + \epsilon_{ij}\) (\(i = 1, \ldots, k; j = 1, \ldots, N_i\)) with \(\epsilon_{ij} \sim NID(0, \sigma^2)\), where the \(\mu_i\) and \(\sigma^2\) are unknown. Define \(\theta_m = \sum_{i=1}^k c_{mi} \mu_i\), where the \(c_{mi}\) are specified constants such that \(\sum_{i=1}^k c_{mi} = 0\) \((m = 1, \ldots, p)\). The \(\theta_m\) represent a family of contrasts among the \(\mu_i\); we suppose that the experimenter is interested in obtaining two-sided interval estimates of the \(\theta_m\) with specified joint confidence coefficient \(1 - \alpha\). In some applications one-sided interval estimates may be more appropriate. This causes no difficulties in principle, but our tables do not apply to that situation.

An interval estimator for \(\theta_m\) is given by

\[
\theta_m = \hat{\theta}_m \pm hs \sqrt{\sum_{i=1}^k c_{mi}^2 / N_i},
\]

where \(\hat{\theta}_m = \sum_{i=1}^k c_{mi} \hat{\mu}_i = \sum_{i=1}^k y_{ij} / N_i\), and \(s^2\) is the usual unbiased estimator of \(\sigma^2\) based on \(v\) df; \(h\) is a constant to be determined to achieve the specified joint confidence coefficient. The random variable involved in determining the confidence interval for \(\theta_m\) \((m = 1, 2, \ldots, p)\) is

\[
T_m = (\hat{\theta}_m - \theta_m) / \sqrt{\sum_{i=1}^k c_{mi}^2 / N_i}\]

We require \(h\) to satisfy

\[
P\left(\bigcap_{m=1}^p |T_m| \leq h\right) = P\left(\max_m T_m^2 \leq h^2\right) = 1 - \alpha.
\]

The distribution of \((T_1, T_2, \ldots, T_p)\) is given by the multivariate Student \(t\) distribution, the joint density function of which was obtained by Dunnett and Sobel (1954) and by Cornish (1954). The correlation coefficient between \(\hat{\theta}_m\) and \(\hat{\theta}_m\) is given by

\[
\rho_{m_1, m_2} = \left[\frac{\sum_{i=1}^k c_{m_1 i} c_{m_2 i} / N_i}{\left(\sum_{i=1}^k c_{m_1 i}^2 / N_i\right)^{1/2}} \left(\sum_{i=1}^k c_{m_2 i}^2 / N_i\right)^{1/2}\right]^{1/2}.
\]

A special case of interest is \(\rho_{m_1, m_2} = \rho\), for all \(m_1 \neq m_2\). For this case, the \(x\) points of the distribution of \(\max_m |T_m|\), denoted by \(h(x, \rho, \alpha, \alpha)\), have been tabulated for selected values of \(\rho\) and \(\alpha\), namely, by Dunnett (1964) for \(\rho = .5\) and \(\alpha = .05, .01\); and by Hahn and Hendrickson (1971) for \(\rho = .0, .2, .4, .5\) and \(\alpha = .10, .05, .01\). (These tables are also given in Miller 1981.) In addition, Krishnaiah and Armitage (1970) tabulated \(h^2\) for \(\rho = .1(.1).9\) and \(\alpha = .05\) and .01.
2.2 Special Case \( p = 0 \): Studentized Maximum Modulus

The special case \( \rho_{m_1, m_2} = p = 0 \) is of particular interest. In this case \( \max_{m_1} | T_{m_1}^p \) is known as the Studentized maximum modulus. It was originally tabulated for \( \alpha = .05 \) by Pillai and Ramachandran (1954). More recently, it has been tabulated by Stoline and Ury (1979) for \( p = k(k - 1)/2, k = 3(1)20 \), and \( \alpha = .2, .1, .05, .01 \) and by Ury, Stoline, and Mitchell (1980) for \( p = k(k - 1)/2, k = 20(2)50(5)80, 90, 100 \) and \( \alpha = .2, .1, .05, .01 \).

In our applications the case \( p = 0 \) arises when \( N_i = N \) and \( \sum_{i=1}^{k} c_{m_1} c_{m_2} = 0 \), viz., for balanced designs and a mutually orthogonal set of contrasts. In such situations the constants \( h = h_{1}(p, 0, \alpha) \) provide the exact specified joint confidence coefficient for the interval estimates of the contrasts. Table 1 of this article gives \( h = h_{1}(p, 0, \alpha) \), correct to two decimal places, for \( \alpha = .2, .1, .05, .01 \) and \( v = 20(2)50(5)80, 90, 100 \) and \( \alpha = .2, .1, .05, .01 \).

Table 1 is a condensation of tables given in Bechhofer and Dunnett 1981, which contain \( h \), correct to four decimal places, for \( \alpha = .20, .10, .05, .01 \). The computations for these tables were carried out by numerical integration of an infinite integral using a 96-point Legendre quadrature formula given in Abramowitz and Stegun 1964, p. 919; the reader is referred to our earlier article for details.

Table 1 is more suitable for our problem than those of Stoline because the latter tabulate \( h \) only for \( p = 3, 6, 10, 15, 21, 28, \) and so on, and hence interpolation on \( p \) would be required. Our tables are limited to \( p \leq 32 \); for values of \( p > 32 \) (which should occur infrequently in practice) the tables of Stoline and Ury (1979) can be used with interpolation.

Finney (1941) considered the distribution of the maximum of several \( F \) ratios arising in an analysis of variance, the numerator mean squares being independent each with the same number of df, denoted by \( 2m \), and the denominators being based on a common mean square. For \( m = \frac{1}{2} \), Finney's max \( F \) is identical with the square of the Studentized maximum modulus, and hence the \( \alpha \) points of Finney's statistic for the special case of \( F \) ratios with a single df in the numerator are given by the squares of the values given in our tables. Nair (1948) provided a table of such \( \alpha \) points for \( \alpha = .05 \) and \( p = 1(1)10 \) and \( v = 10, 12, 15, 20, 30, 60, \infty \); he used a data set of Wishart (1938) arising from a \( 2^3 \) experiment to illustrate the proper use of max \( F \) to control the error probability when multiple testing of effects is carried out.

It should be noted that the constants \( h_{1}(p, 0, \alpha) \) can also be used as conservative approximations (i.e., the joint confidence coefficient achieved will exceed the nominal \( 1 - \alpha \)) if the design is unbalanced, and also if a set of nonorthogonal contrasts is specified. In the latter case, use of the Bonferroni statistic has often been recommended (see, e.g., Miller 1981, p. 68), or the slightly improved bounds obtained by the Dunn–Šidák method (Dunn 1974), which makes use of Šidák's (1967) multiplicative inequality. However, use of the Studentized maximum modulus produces shorter intervals than either of these procedures; this can be seen by comparing the entries in our tables with those given by Bailey (1977) for the Bonferroni method and with those given by Games (1977) for the Dunn–Šidák method. (The conservativeness of the values of \( h \) given by the Studentized maximum modulus follows from the uncorrelated \( t \) inequality of Šidák 1967.) We have prepared a short table (Table 2) containing corresponding values of the Bonferroni \( t \) and the Studentized maximum modulus for selected \( v, \alpha, \) and \( p \); these are given in three decimal places, the entries for the Studentized maximum modulus being obtained by rounding the four decimal-place entries in Bechhofer and Dunnett (1981). A study of the differences between the corresponding entries shows that (a) for fixed \( \alpha \) and \( v \) the difference increases with increasing \( p \), (b) for fixed \( p \) and \( v \) the difference increases with decreasing \( \alpha \) for small \( v \) but decreases with decreasing \( \alpha \) for sufficiently large \( v \), (c) for fixed \( p \) and \( \alpha \) the difference decreases with increasing \( v \). For \( v = 20 \) the maximum difference for \( p \leq 30, \alpha \leq .10 \) is only \( 3.331 - 3.235 = .096 \) and the corresponding increase in interval width is only \( .096/3.235 \times 100 = 3 \) percent; hence, if \( v \) is sufficiently large the Bonferroni \( t \) yields essentially the same intervals as does the Studentized maximum modulus, although for small \( v \) and large \( p \) the difference can be substantial.

3. ILLUSTRATIONS OF USES OF THE TABLES

In this section we consider several types of planned experiments in which the pertinent inferences are made on the basis of orthogonal contrasts among the treatment means. We shall point out some of the issues involved and indicate how the multiple comparisons procedures used with the appropriate constants in our tables can control the experimentwise error rates for such inferences.

3.1 Experiments Involving a Single Qualitative Factor

Example 1. A Five-Level Experiment. Bennett and Franklin (1954, Sec. 7.34) consider an experiment involving five different methods of analyzing the concentration of iron in a standard solution. Two methods included agitation and three methods did not; four analyses were made with each method. The orthogonal contrasts under consideration (see their Table 7.9) were \( \{c_{m_1}, c_{m_2}, \ldots, c_{m_3}\} = (3, 3, -2, -2, -2, 1, -1, 0, 0, 0, 0, 0, 2, -1, -1, -1, -1) \) for \( m = 1, 2, 3, 4, \) respectively. Here \( p = 4, v = 15, \) and
<table>
<thead>
<tr>
<th>p</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>.1</td>
<td>.1</td>
<td>.2</td>
<td>.3</td>
<td>.4</td>
<td>.5</td>
<td>.6</td>
<td>.7</td>
<td>.8</td>
<td>.9</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>.15</td>
<td>.19</td>
<td>.22</td>
<td>.24</td>
<td>.26</td>
<td>.28</td>
<td>.3</td>
<td>.3</td>
<td>.31</td>
<td>.3</td>
<td>.3</td>
<td>.3</td>
<td>.3</td>
<td>.3</td>
<td>.3</td>
<td>.3</td>
</tr>
<tr>
<td>.15</td>
<td>.19</td>
<td>.22</td>
<td>.24</td>
<td>.26</td>
<td>.28</td>
<td>.3</td>
<td>.3</td>
<td>.31</td>
<td>.3</td>
<td>.3</td>
<td>.3</td>
<td>.3</td>
<td>.3</td>
<td>.3</td>
<td>.3</td>
</tr>
<tr>
<td>.15</td>
<td>.19</td>
<td>.22</td>
<td>.24</td>
<td>.26</td>
<td>.28</td>
<td>.3</td>
<td>.3</td>
<td>.31</td>
<td>.3</td>
<td>.3</td>
<td>.3</td>
<td>.3</td>
<td>.3</td>
<td>.3</td>
<td>.3</td>
</tr>
<tr>
<td>.15</td>
<td>.19</td>
<td>.22</td>
<td>.24</td>
<td>.26</td>
<td>.28</td>
<td>.3</td>
<td>.3</td>
<td>.31</td>
<td>.3</td>
<td>.3</td>
<td>.3</td>
<td>.3</td>
<td>.3</td>
<td>.3</td>
<td>.3</td>
</tr>
</tbody>
</table>

Table 1. Studentized Maximum Modulus for $p = 2$ and 16 Contrasts
<table>
<thead>
<tr>
<th>p</th>
<th>17</th>
<th>18</th>
<th>19</th>
<th>20</th>
<th>21</th>
<th>22</th>
<th>23</th>
<th>24</th>
<th>25</th>
<th>26</th>
<th>27</th>
<th>28</th>
<th>29</th>
<th>30</th>
<th>31</th>
<th>32</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>5.46</td>
<td>5.50</td>
<td>5.54</td>
<td>5.58</td>
<td>5.62</td>
<td>5.66</td>
<td>5.70</td>
<td>5.74</td>
<td>5.78</td>
<td>5.82</td>
<td>5.86</td>
<td>5.90</td>
<td>5.93</td>
<td>5.97</td>
<td>6.00</td>
<td>6.03</td>
</tr>
<tr>
<td>4</td>
<td>2.47</td>
<td>2.51</td>
<td>2.55</td>
<td>2.59</td>
<td>2.63</td>
<td>2.67</td>
<td>2.71</td>
<td>2.75</td>
<td>2.79</td>
<td>2.83</td>
<td>2.87</td>
<td>2.90</td>
<td>2.93</td>
<td>2.96</td>
<td>2.99</td>
<td>3.02</td>
</tr>
</tbody>
</table>

Table 1. (continued) Studentized Maximum Modulus for p = 17-32 Contrasts
Table 2. Values of Bonferroni t (upper entry) and Studentized Maximum Modulus (lower entry) for Selected \( v \), \( \alpha \), and \( p \)

<table>
<thead>
<tr>
<th>( v )</th>
<th>( \alpha )</th>
<th>( p = 2 )</th>
<th>( p = 10 )</th>
<th>( p = 30 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>0.10</td>
<td>4.303</td>
<td>9.925</td>
<td>17.277</td>
</tr>
<tr>
<td>10</td>
<td>0.05</td>
<td>6.205</td>
<td>14.089</td>
<td>24.404</td>
</tr>
<tr>
<td>10</td>
<td>0.01</td>
<td>14.089</td>
<td>31.599</td>
<td>54.759</td>
</tr>
</tbody>
</table>

If the number of analyses made with each method were not all equal, some (and possibly all) of the contrasts under consideration would not be orthogonal. Then the use of our \( h \) values yields a confidence coefficient greater than \( 1 - \alpha \). (See Šidák 1967, Eq. (8).)

3.2 2\textsuperscript{a} Factorial Experiments

Example 2: A 2\textsuperscript{a} Experiment. Cochran and Cox (1957, Sec. 5.24a) analyze the yields obtained in a 2\textsuperscript{a} experiment with four fertilizers \((m = \text{manure}, n = \text{nitrogen}, p = \text{phosphorus}, k = \text{potassium})\), each at two levels, conducted in four randomized complete blocks; the experiment was carried out to study the effects of these fertilizers on the yield of grass. The 64 yields per plot (total over six harvests, per three-meter row) were used to compute the effect means, which were given as \( M = 18.0, N = 21.3, P = 24.1, MN = 3.2, MP = -0.3, MK = -7.5, NP = 5.5, NK = 10.9, PK = 3.2, MNK = -1.4, MNK = -8.5, MPK = 0.8, NPK = 0.5, MNPK = -1.6. The estimated standard error of an effect mean was computed to be

\[
\sqrt{\frac{s^2}{2^a-4}} = \sqrt{\frac{90.5}{4(4)}} = 2.38,
\]

where \( s^2 = 90.5 \) is the error mean square based on \( v = 45 \), \( r = 4 \) is the number of replications of each treatment combination, and \( n - 4 \) is the number of factors. For \( v = 45 \) the Student t values are 2.69 and 2.01 for \( \alpha = 0.01 \) and \( \alpha = 0.05 \), respectively. The effect means ± \( t_{24}(2.38) \) are exhibited in Columns 2 and 3 of Table 3 for \( \alpha = 0.01 \) and \( \alpha = 0.05 \), respectively. Cochran and Cox give analogous information in the lower half of their Table 5.1a, where they indicate effects that are statistically significant at the 1 percent (**) and 5 percent (*) levels, namely \((M, N, K, MK, NK, MNK)\) and \((P)\), respectively.

To control the experimentwise error rate, the corresponding \( h \) values from our Table 1 are \( h_{15}(4, 0, .01) = 3.65 \) and \( h_{15}(15, 0, .05) = 3.08 \) for \( \alpha = 0.01 \) and \( \alpha = 0.05 \), respectively. The effect means ± \( h_{15}(4, 0, \alpha)(2.38) \) are exhibited in Columns 4 and 5 of Table 3 for \( \alpha = 0.01 \) and \( \alpha = 0.05 \), respectively. Cochran and Cox give analogous information in the lower half of their Table 5.1a, where they indicate effects that are statistically significant at the 1 percent (**) and 5 percent (*) levels, namely \((M, N, K, MK, NK, MNK)\) and \((P)\), respectively.

The experimenter must decide whether the per contrast or the experimentwise error rate is more pertinent in the particular experiment at hand. In the extreme case of a single significant contrast, the experimenter using per contrast error rates would be left feeling unsure whether the effect was a real one, given that the experiment has provided multiple op-
opportunities for one of the effects to be significant. On the other hand, in an experiment with several significant contrasts, as in the present example, strict use of the experimentwise error rate procedure makes it more difficult for the smaller effects to be declared significant after the larger ones have been identified.

We assumed in the previous analysis that a priori the experimenter was interested in controlling the experimental error rate for all 15 contrasts. If, a priori, this experimenter had been interested in only the four main effects and the six two-factor interactions (and made that decision without being influenced by the data), then \( p = 10 \), \( v = 45 \), and the appropriate \( h \) value for \( \alpha = 0.01 \) and \( \alpha = 0.05 \) would be 3.51 and 2.93, respectively; now MK is still significant at the 5 percent level and \( P \) is still not significant at the 5 percent level, while the status of MNK (as well as MNP, MPK, and MNPK) in terms of possible significance would be unknown. If, after looking at the data, the experimenter decided that only \( M \), \( N \), \( K \), \( NK \), and \( MNK \) were of interest, then the original factor \( h_{\text{ext}}(15, 0, z) \) must still be used in reporting the final results.

Effectively, what has been done here is "data snooping" in the sense of Scheffé (1959 p. 80) and that privilege must be paid for by using the larger \( h \) value if statistically legitimate confidence statements with experimentwise control over the error rate are desired. In this situation the inference must be limited to a particular set of orthogonal contrasts specified in advance.

It thus is clear that in a complete factorial experiments where \( n \) is large, it is to the experimenter's advantage if he can specify a priori which contrasts are and/or are not of interest; analogous considerations hold for fractional factorial experiments. In certain types of experiments it is easy to identify certain contrasts that are not of interest. We consider such a problem in Example 3.

Example 3: A 2^3 Experiment With Two Classification Factors. Consider a three-factor experiment, each factor at two levels, where the factors are diet (Diet 1 vs. Diet 2), sex (male vs. female), and age (old vs. young). The purpose of the experiment is to study the effect of change in diet on gain in weight. Here the treatment factor of interest is diet while sex and age are classification variables. Thus, denoting the main effect of diet, sex, and age by \( A \), \( B \), and \( C \), respectively, and analogously for their interactions, the experimenter would be interested in the \( p = 4 \) orthogonal contrasts associated with \( A \), \( AB \), \( AC \), and \( ABC \) rather than in all seven orthogonal contrasts. Similarly, in a complete factorial experiment, two of which are classification factors, the experimenter would be interested in at most the \( p = 12 \) orthogonal contrasts associated with \( A \), \( AB \), \( AC \), \( AD \), \( ACD \), \( B \), \( BC \), \( BD \), \( BCD \), \( ABC \), \( ABD \), \( ABCD \). See Cox (1958, Examples 6.3 and 6.4) for a discussion of treatment factors and classification factors.

3.3 3^a Factorial Experiments, All Factors Quantitative

Example 4: A 3^3 Experiment. Davies (1978, pp. 332-336) reports the results of a 3^3 experiment, each factor quantitative and equally spaced, all treatment combinations replicated twice. The variable under study is the yield of a chemical process, the three factors being: (a) \( C \), the concentration of an inorganic material (A) in the free water present in the reaction mixture, (b) \( V \), the volume of free water present in the reaction mixture, and (c) \( N \), the amount of a second inorganic material (B) in the reaction mixture. The model fitted to the data used orthogonal polynomials, and the total df for treatments was partitioned into 18 individual df associated with the 6 main effects \( (C_L, C_Q, V_L, V_Q, N_L, N_Q) \) and the 12 two-factor interactions \( (C_L \times V_L, C_Q \times V_Q, C_L \times V_Q, C_Q \times V_L, C_L \times N_L, C_Q \times N_Q, C_L \times Q \times N_Q, C_Q \times N_L, V_L \times D, V_Q \times N_Q, N_Q \times N_L, V_Q \times N_L, V_L \times \cdot N_Q) \); the remaining eight degrees of freedom representing the three-factor interactions were pooled. Here the subscripts \( L \) and \( Q \) represent linear and quadratic, respec-

---

**Table 3. A 2^4 Complete Factorial Experiment**

\((p = 15, v = 45)^1\)

<table>
<thead>
<tr>
<th>Effect</th>
<th>(a = 0.01)</th>
<th>(a = 0.05)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Error rate (h) per contrast</td>
<td>Error rate (h) per experiment</td>
<td></td>
</tr>
<tr>
<td>(i = 5, f = 5)</td>
<td>2.69</td>
<td>2.01</td>
</tr>
</tbody>
</table>

1. From Abramowitz and Stegun (1964), Table 5.1.

2. Standard error of effect mean is \( \sigma/\sqrt{2m+1} = \sigma/\sqrt{2(35)} = 2.38 \).

The entry in each cell is the critical value of the test statistic, \( h_{ij}(2.38) \).

The intervals indicated by (\( \ast \)) or (\( \ast\ast \)) do not cover zero.
tively. There were 27 df associated with the error mean square.

For \( v = 27 \) the Student \( t \) values are 2.77 and 2.05, for \( \alpha = .01 \) and \( \alpha = .05 \), respectively. Using these values (actually the corresponding \( F_{27} \) values were used) Davies reported the eight effects (\( C_L, C_Q, V_L, V_Q, C_Q \times V_L, C_Q \times V_Q, C_E \times V_Q, C_Q \times N_Q \)) as being statistically significant at the 1 percent level and the two effects (\( N_Q, C_L \times N_Q \)) significant at the 10 percent level. (Note: We are reporting the results here as tests of significance to conform with Davies, but we would have preferred to present our results as interval estimates as in our Table 3.)

To control the experimentwise error rate the corresponding \( h \) values from our Table 1 are \( h_{27}(18, 0, .01) = 3.90 \), \( h_{27}(18, 0, .05) = 3.26 \), and \( h_{27}(18, 0, .10) = 2.96 \) for \( \alpha = .01 \), \( \alpha = .05 \), and \( \alpha = .10 \), respectively. Thus, if the experimenter wished to control the experimentwise error rate for the 18 orthogonal contrasts of interest, he would assert that the three effects (\( C_L, C_Q, C_L \times V_Q \)) are significant at the 1 percent level, the three effects (\( V_L, V_Q, C_Q \times N_Q \)) are significant at the 5 percent level, the one effect (\( C_Q \times V_L \)) is significant at the 10 percent level, and the remaining 11 effects are not significant at the 10 percent level. (The experimenter had decided a priori that the eight effects associated with the three-factor interactions were not of interest, and hence this total sum of squares based on eight df was not partitioned into the eight individual one df sums of squares associated with each of the remaining relevant orthogonal contrasts.)

Thus the same considerations arise in the analysis of this experiment as arose in the analysis of the 24 experiment of Example 2. And the same caveats hold here as well.

### 3.4 An Application to Biological Assay

The purpose of biological assay is to estimate the potency ratio, \( \rho \), of two biological preparations having dose-response curves that can be represented by the same form of regression function and that differ only in the factor \( \rho \) in the dose scale. (We use the symbol \( \rho \) here as in Finney’s, 1978, p. 41.) A common situation is one in which the response scale is linear in log dose. In this case parallel straight lines can be fitted to the two sets of data; an estimate of log \( \rho \) is then given by the horizontal distance between them. A useful experimental design for such situations is the so-called symmetric (\( k, k \))-point design in which \( k \) dose levels equally spaced on a log scale are used for each preparation (the \( k \) dose levels being chosen so that expected responses from corresponding dose levels are equal); \( n \) observations are taken at each of the \( 2k \) design points, the usual values for \( k \) being 2, 3, or 4. See Finney (1978, p. 105).

When the data from a (\( k, k \)) bioassay are analyzed, the sum of squares between treatments has \( 2k - 1 \) df, which can be separated into \( 2k - 1 \) meaningful orthogonal components. See Finney (1978, pp. 105–109) for an example with \( k = 3 \), where the five orthogonal contrasts are denoted by \( L_p \) (preparations), \( L_1 \) (average linear regression), \( L_2 \) (parallelism), \( L_3 \) (average quadratic regression), \( L_4 \) (difference between quadratics). (Alternatively \( L_2 \) and \( L_4 \) could be defined to be the curvatures for the standard and test preparations, respectively.) The first two enter into the calculation of the estimated relative potency while the remaining three are used to test the validity of the assay. However, significance tests at level \( \alpha = .05 \) for each of these three will result in an error rate for the assay approaching \( 1 - (1 - \alpha)^3 \). If it is desired to control the experimentwise error rate at a specified level \( \alpha \), then the constants \( h \) tabulated in our article can be used.

To apply our constants to Finney’s example, the largest of the three mean squares for the validity contrasts, which here is \( L_2 \) having a value of .001606 (see Finney’s table 5.2.2), is expressed as a ratio to the error mean square based on 30 df; this ratio then is compared for \( \alpha = .05 \) with \( h_{30}(3, 0, .05) = (2.52)^2 = 6.35 \) instead of referring it to tables of \( F_{30} \). In this experiment the ratio is only .52, a clearly nonsignificant result.

In Finney’s description, the contrast \( L_p \) is also considered as a test of a type of assay validity, and as such it could be included with the other three to form a set of four simultaneous tests involving orthogonal contrasts. Actually, \( L_p \) provides a measure of how successful the experimenter has been in choosing comparable dose levels of the two preparations, and a significant value provides a signal to the experimenter that he might be comparing the two preparations at different portions of the dose-response curve rather than necessarily invalidating the assay. Thus it might be preferable to consider \( L_p \) separately from the other three validity contrasts, as illustrated in the preceding paragraph.

Similar considerations apply with more than three dose levels of each preparation. For example, in a (4, 4) bioassay there would be five orthogonal contrasts in addition to \( L_p \) and \( L_1 \). If linearity is assumed then these provide five separate tests for assay validity, which can be tested by using \( h_{4}(5, 0, a) \) from our tables to achieve an experimentwise error \( a \) for the validity tests. On the other hand, if a quadratic dose-response curve is assumed, two of the contrasts enter into the calculation of the estimated potency, as described by Finney (1978, p. 122), leaving a set of three orthogonal contrasts to test the validity of the bioassay.

It should be pointed out that, although we have described the problem of testing validity in bioassay as a problem in multiple significance testing of a set of
orthogonal contrasts, it might be more complex than this in practice. For example, two moderately large but nonsignificant contrasts might be as worrisome as one significant contrast. In this case, one might wish to allow for the possibility that more general contrasts should also be tested. Also, confidence intervals on the contrasts would probably be more appropriate since the pertinent question is really whether the population contrasts are large enough to invalidate the assay rather than whether they are zero. Furthermore, many laboratories are geared to perform bioassays routinely so that they would have a considerable amount of past experience available that should be used in judging the validity of the assay.

4. CONCLUDING REMARKS

In this article we have been concerned with experiments in which the pertinent inferences are made on the basis of orthogonal contrasts among treatment means. The experiments may be single-factor or multifactor and may involve qualitative and/or quantitative variables. Several examples of such experiments were described in Section 3. For each example we showed how an appropriate multiple comparisons procedure could be used to estimate simultaneously a set of meaningful orthogonal contrasts, and to do so in such a way that the overall error probability, or more precisely the joint confidence coefficient, is controlled at some prespecified level. This was achieved by using the upper percentage point of the distribution of the Studentized maximum modulus statistic. A table of these percentage points was prepared for this article and is given as Table 1.

If the treatment means from which the contrasts are calculated are based on unequal numbers of observations, then the contrasts are nonorthogonal. Nonorthogonal contrasts can also arise if the population contrasts that are of particular interest to the experimenter are themselves nonorthogonal. In both situations the constants that we provided can still be used, the only effect being that they will lead to conservative intervals; that is, the actual achieved joint confidence coefficient will be greater than the nominal value that was specified.

The constants can also be used in situations in which the Bonferroni t method has sometimes been advocated and results in an improvement over the latter method since the Studentized maximum modulus values are always smaller than the corresponding Bonferroni t values; however, the difference between the two may be small, particularly for large degrees of freedom.

In any particular experiment the experimenter must decide whether or not the planned inferences require a simultaneous inference procedure, that is, a multiple-comparisons procedure. We caution that when several tests are performed on the same data set, the probability of making a wrong inference is increased unless an appropriate multiple comparisons procedure is used. Moreover, if the experimenter’s inferences involve a set of orthogonal contrasts, it is advantageous to specify in advance the smallest possible number of such contrasts to which to restrict attention.

5. ACKNOWLEDGMENTS

The authors are indebted to the editor and a referee for many constructive suggestions. The research of the first author was supported by U.S. Army Research Office-Durham Contract DAAG29-80-C-0036 and Office of Naval Research Contract N00014-75-C-0586 at Cornell University, and that of the second author by the Natural Sciences and Engineering Research Council of Canada at McMaster University.

[Received March 1981. Revised March 1982.]

REFERENCES


