STATISTICS IN MEDICINE, VOL. 10, 939–947 (1991)

STEP-DOWN MULTIPLE TESTS FOR COMPARING TREATMENTS WITH A CONTROL IN UNBALANCED ONE-WAY LAYOUTS

CHARLES W. DUNNETT

Department of Clinical Epidemiology & Biostatistics and Department of Mathematics & Statistics, McMaster University, Hamilton, Ontario L8S 4K1, Canada

AND

AJIT C. TAMHANE

Department of Statistics and Department of Industrial Engineering & Management Sciences, Northwestern University, Evanston, Illinois 60208, U.S.A.

SUMMARY

We show how a well-known multiple step-down significance testing procedure for comparing treatments with a control in balanced one-way layouts can be applied in unbalanced layouts (unequal sample sizes for the treatments). The method we describe has the advantage that it provides p-values, for each treatment versus control comparison, that take account of the multiple step-down testing nature of the procedure. These *joint p-values* can be used with any value of α, the fixed type I familywise error rate bound, that may be specified by the investigator. To determine the p-values, it is necessary to compute a multivariate Student t integral, for which a computer program is available. This procedure is more powerful than the step-down Bonferroni procedure of Holm[1] and the single-step procedure of Dunnett.[2] An example from the pharmaceutical literature is used to illustrate the procedure.

1. INTRODUCTION

In many studies, comparisons between treatments and a specified treatment may be of interest for the purpose of determining which treatments are different from the specified treatment and/or estimating the differences from the specified treatment. For example, the specified treatment may be a placebo control or an accepted standard therapy in a medical trial with which one or more alternative therapies are compared. Another example is a drug trial where a new drug is compared with one or more reference standard drugs.

The purpose of such studies is usually to estimate differences between the various treatments and the specified treatment with respect to certain response variables which may be critical for determining the efficacy or safety of the treatments. Sometimes tests of significance are performed on the treatment versus control differences for a particular response variable to test whether they exceed or differ from some specified threshold value (which may be zero or some $\delta \neq 0$). If the study contains more than one test treatment to be compared with the control, then the experimenter may need to take account of the multiple testing being performed in order to provide sufficiently convincing evidence that a particular observed difference is indicative of a real effect rather than being a result of the multiplicity.

0277–6715/91/060939–09$05.00

The first to recognize the need for allowing for multiple testing when testing the significance of differences between treatments and a control was Roessler.[3] He proposed that the customary Student t-tests be performed at a level $1-(1-\alpha)^{1/k}$ instead of α when comparing k treatments with a control; the implication was that this will control the probability of making any type I error at approximately α. This probability is the so-called *familywise error rate*, FWE, and we refer to its nominal value, α, as the *joint significance level*. (Later it was shown[4,5] that α was an upper bound for the actual FWE.) Dunnett[2] developed an exact method for maintaining the FWE at α based on the multivariate t-distribution. This method was formally described in terms of one-sided or two-sided simultaneous confidence interval estimates of the k treatment versus control differences, with specified joint confidence coefficient, $1-\alpha$. Using this method each individual difference could be tested at a joint significance level α by observing whether zero (or any other specified value δ) was included in the corresponding confidence interval.

However, if only a significance testing procedure was required, it was clear that the power of the tests could be increased further by performing them in a step-down manner similar to the well-known Newman–Keuls test for pairwise treatment comparisons. The first to point this out was Miller[6] (see pp. 78 and 85–86). Naik[7] proposed the same procedure in a different context. The reason for the increase in power is that the successive tests use a decreasing sequence of critical values based on m-variate Student t distributions with $m=k,\ k-1,\ldots,1$, these critical values being less than the critical value required for the simultaneous confidence intervals except when $m=k$. Marcus *et al.*[8] showed, using their closure principle, that this step-down testing method does, in fact, maintain the FWE at or below the desired level α. Although they considered a balanced one-way layout (that is, equal sample sizes for all the treatments), the procedure can be readily extended and applied when the test treatments have a common sample size which is different from that of the control, since the associated m-variate t distributions are equicorrelated and tables of critical points are available.[9] Marcus *et al.*'s[8] proof applies also to completely unbalanced one-way layouts, but the application of the procedure in this case is hampered by the difficulty of computing exact critical constants at each step of testing, since these critical constants require the use of multivariate t-distributions with unequal correlations to compute their values.

A conservative solution to this difficulty is provided by the sequentially rejective Bonferroni procedure of Holm.[1] This is a step-down testing procedure which uses the Bonferroni upper bounds on the exact multivariate t critical points, the Bonferroni bounds being based on the univariate t distribution. In the present paper we show how, by using a computer program by Dunnett[10] which calculates probability integrals of the multivariate normal distribution having a so-called product correlation structure, the Bonferroni approximation can be dispensed with and an exact application of the step-down test procedure in unbalanced one-way layouts becomes possible.

In fact, our method computes 'joint' p-values associated with the observed treatment versus control mean differences, taking into account the multiple step-down testing nature of the procedure. Here, as in the case of a single test (the p-value for which may be referred to as a 'marginal' p-value), the *joint p-value* for any observed difference is defined to be the smallest joint level of significance α at which that difference would be declared statistically significant. (From now on, we will drop the prefix 'joint' when referring to p-values and α if there is no possibility of ambiguity.) These p-values are more informative than merely stating whether the observed difference is significant at specified level α. It should be noted that the concept of p-values has not been used much in the multiple comparisons literature, a recent exception being Westfall and Young.[11]

The outline of the paper is as follows: Section 2 describes the step-down test procedure, both in its classical version based on critical constants for specified α and in its p-value version as proposed here. Two approximations to the classical version, both of which can be applied without access to a computer program, are also described; one is the Bonferroni approximation due to Holm,[1] while the other is based on a close approximation to the required multivariate Student's t critical points. Section 3 gives the details of the computation of the multivariate t probability integral. Section 4 describes an illustrative example followed by discussion in Section 5.

2. DESCRIPTION OF THE STEP-DOWN TESTING PROCEDURES

We use the subscripts $1, 2, \ldots, k$ to denote $k \geqslant 2$ test treatments and the subscript 0 to denote the control or otherwise specified treatment. Denote by n_i the sample size for the ith treatment $(0 \leqslant i \leqslant k)$. We assume the standard one-way fixed effects model, that is, normality with unknown treatment means μ_i and common unknown error variance σ^2. Thus the sample means $\bar{X}_i$ are mutually independent $N(\mu_i, \sigma^2/n_i)$ random variables distributed independently of the pooled sample variance S^2. The latter is a $\sigma^2\chi_\nu^2/\nu$ random variable based on $\nu = \sum n_i - (k+1)$ degrees of freedom (d.f.).

Consider the problem of testing the hypotheses H_{0i}: $\mu_i - \mu_0 \leqslant \delta$ against upper one-sided alternatives H_{1i}: $\mu_i - \mu_0 > \delta$ $(1 \leqslant i \leqslant k)$ where δ is a specified constant. Denote the observed values of the $\bar{X}_i$ and S^2 by the corresponding lower case letters $\bar{x}_i$ and s^2, respectively $(0 \leqslant i \leqslant k)$. The Student t-statistics for testing the H_{0i} are given by

$$t_i = \frac{(\bar{x}_i - \bar{x}_0 - \delta)}{s\sqrt{(1/n_i + 1/n_0)}} \qquad (1 \leqslant i \leqslant k). \tag{1}$$

If we let

$$T_i = \frac{\{\bar{X}_i - \bar{X}_0 - (\mu_i - \mu_0)\}}{S\sqrt{(1/n_i + 1/n_0)}} \qquad (1 \leqslant i \leqslant k) \tag{2}$$

then $T_1, T_2, \ldots, T_k$ have a joint k-variate central t-distribution with ν d.f. and associated correlation coefficients given by $\rho_{ij} = \lambda_i\lambda_j$ where $\lambda_i = 1/\sqrt{(1 + n_0/n_i)}$ $(1 \leqslant i \neq j \leqslant k)$. The resulting correlation matrix, say R_k, is said to possess a product correlation structure (Hochberg and Tamhane,[12] p. 365; see also Curnow and Dunnett[13]). Let $t^{(\alpha)}_{m,\nu,R_m}$ be the upper α point of the maximum of $T_1, \ldots, T_m$, which have an m-variate t-distribution with ν d.f. and associated correlation matrix R_m $(m = 1, \ldots, k)$. Note that R_m is the submatrix of R_k formed by its first m rows and m columns.

We shall describe the step-down test procedure first in its classical version, based on critical constants for a specified α. From now on we will assume that the treatments and the associated hypotheses are relabelled so that $t_1 \leqslant t_2 \leqslant \ldots \leqslant t_k$. (However, the random variables T_i's are not assumed to be ordered.) The step-down test procedure rejects H_{0m} iff $H_{0,m+1}, \ldots, H_{0k}$ have been rejected and $t_m > t^{(\alpha)}_{m,\nu,R_m}$; if H_{0m} is accepted then $H_{01}, \ldots, H_{0,m-1}$ are accepted by implication without actually testing them. It can be shown that this procedure controls the type I FWE at level α (see Hochberg and Tamhane,[11] pp. 54–56). It is also easy to see that the critical constants $t^{(\alpha)}_{m,\nu,R_m}$ are monotonically increasing in m. Dunnett's[2] single-step procedure uses the largest of these critical constants, namely $t^{(\alpha)}_{k,\nu,R_k}$ for testing all the hypotheses (without regard to their order), and hence it is less powerful than its step-down counterpart.

Next we describe the p-value version of this step-down procedure. First, compute

$$p'_m = P\{\text{at least one } T_i \geqslant t_m,\ i = 1, \ldots, m\}$$
$$= 1 - P\{T_i < t_m,\ i = 1, \ldots, m\} \qquad (m = 1, \ldots, k). \tag{3}$$

Then define the p-value for H_{0m} (denoted by p_m) to be

$$p_m = \begin{cases} p'_k & \text{for } m = k \\ \max(p'_m, p_{m+1}) & \text{for } m = 1, \ldots, k-1. \end{cases} \tag{4}$$

An alternative way of writing (4) is $p_m = \max(p'_m, p'_{m+1}, \ldots, p'_k)$.

Once these p-values are determined, hypothesis tests can be conducted at any fixed specified level α, if desired, by comparing any p_m with α and rejecting H_{0m} if $p_m \leqslant \alpha$ $(m = 1, \ldots, k)$. In other situations it may be more useful to simply report the p-values and perhaps use them as inverse measures of the strength of evidence in favour of the H_{1m}.

Note that the p-values are monotonically ordered, that is, $p_1 \geqslant p_2 \geqslant \ldots \geqslant p_k$, whereas the p'-values are not always so ordered. (This is in contrast to the critical constants $t^{(\alpha)}_{m,\nu,R_m}$ which are always ordered.) Thus if $p_m > \alpha$ and hence H_{0m} is accepted, then monotonicity ensures acceptance also of $H_{01}, \ldots, H_{0,m-1}$. Therefore the results of both versions of the step-down test procedure will be in accord for whatever value is specified for α.

A computer program is needed to implement either version of this step-down procedure; in the next section we describe a computer program to implement the p-value version. If a user prefers not to have to depend upon a computer program, then the classical (fixed α) version can still be implemented by using available tables and making certain approximations. One approximation is to replace the off-diagonal entries in the correlation matrix R_m by their arithmetic average. Then the required critical constants can be approximated by the corresponding critical constants from equicorrelated multivariate t-distributions, for which extensive tables have been given by Bechhofer and Dunnett.[9] A numerical study of this approximation by Dunnett[14] (see also Hochberg and Tamhane,[11] p. 145), shows that it is generally more accurate than the approximation proposed by Dutt *et al.*[15] and also it is conservative. However, there is no analytical proof that this approximation is always conservative.

A more conservative approach but one that is easier to apply is to use the Bonferroni approximation. This involves replacing the critical constants $t^{(\alpha)}_{m,\nu,R_m}$ by $t^{(\alpha/m)}_{\nu}$, which can be obtained from Bailey's[16] univariate Student t tables constructed especially for using the Bonferroni approximation. An equivalent way of applying this procedure as given by Holm[1] is as follows: let $p''_m = P\{T_m \geqslant t_m\}$, which is the 'marginal' p-value of t_m if that were the only statistic under test $(m = 1, \ldots, k)$. Then mp''_m is the Bonferroni upper bound on p_m. Reject H_{0m} iff $H_{0,m+1}, \ldots, H_{0k}$ are rejected and $mp''_m \leqslant \alpha$; if H_{0m} is accepted then accept $H_{01}, \ldots, H_{0,m-1}$ by implication without actually testing them. This procedure is of step-down type, which is required because the Bonferroni bounds mp''_m are not necessarily ordered although the p''_m are ordered. Slightly less conservative modifications of the Bonferroni approximations are given by Hochberg and Tamhane,[11] p. 148, and by Holland and Copenhaver.[17] Both of these approximations involve replacing all of the ρ_{ij} by zeros.

The above discussion can be extended in an obvious way to the two-sided testing problem, that is, $H_{0i}: \mu_i - \mu_0 = \delta$ versus $H_{1i}: \mu_i - \mu_0 \neq \delta$ $(1 \leqslant i \leqslant k)$. In this case, the tests are based on the absolute values of the Student t-statistics in (1), which are ordered and relabelled so that $|t_1| \leqslant |t_2| \leqslant \ldots \leqslant |t_k|$. Thus, analogous to (3), we have

$$p'_m = 1 - P\{-|t_m| < T_i < |t_m|, i = 1, \ldots, m\} \qquad (m = 1, \ldots, k) \tag{5}$$

and the p-values are defined as before by (4).

3. COMPUTATION OF THE PROBABILITY INTEGRALS

In order to implement the step-down procedure defined in the previous section, it is necessary to be able to compute the probabilities (3) or (5) depending on whether one-sided or two-sided tests are required. These probabilities are of the following general form:

$$P\{a_i < T_i < b_i, i = 1, \ldots, m\} \qquad (m = 1, \ldots, k). \tag{6}$$

For the one-sided problem $a_i = -\infty$ and $b_i = t_m$, while for the two-sided problem $a_i = -|t_m|$ and $b_i = |t_m|$, where the t_m are the values of the Student t-statistics for the treatment versus control comparisons calculated from the experimental data according to (1). (However, recall that the labelling is done differently for one-sided and two-sided testing problems.) If we write $T_i = Z_i/U$ $(1 \leqslant i \leqslant k)$ where $Z_1, Z_2, \ldots, Z_k$ are standard normal random variables with $\text{corr}(Z_i, Z_j) = \rho_{ij} = \lambda_i\lambda_j$ $(\lambda_i = 1/\sqrt{\{1 + n_0/n_i\}})$ and U is distributed as a $\sqrt{(\chi_\nu^2/\nu)}$ random variable independent of the Z_i, then (6) can be written (for a derivation, see Hochberg and Tamhane;[12] Appendix 3) as

$$\int_0^\infty \int_{-\infty}^\infty \prod_{i=1}^m \left\{ \Phi\left[\frac{\lambda_i z + b_i u}{\sqrt{(1-\lambda_i^2)}}\right] - \Phi\left[\frac{\lambda_i z + a_i u}{\sqrt{(1-\lambda_i^2)}}\right] \right\} d\Phi(z)\, f_\nu(u) du \tag{7}$$

where $\Phi(\cdot)$ is the standard normal CDF and $f_\nu(\cdot)$ is the PDF of the $\sqrt{(\chi_\nu^2/\nu)}$ random variable. The inner integral is a multivariate normal probability which can be evaluated using Dunnett's[10] Fortran algorithm. The outer integral can be evaluated by numerical integration using this algorithm as a subroutine. A copy of the program to evaluate the outer integral (7) can be obtained from the authors. This program, which uses the numerical integration routine QDAGI from the IMSL subroutine library, was used to compute the multivariate Student probabilities for the example in the next section. (The average computing time when $m = 5$ and an accuracy of 10^{-4} was specified was 52 seconds using a VAX 8600 computer.)

4. ILLUSTRATIVE EXAMPLE

In this section, we illustrate the step-down procedure using an example from a recent investigation by Bedotto *et al.*[18] This was a study in rats to elucidate the manner in which thyroid hormone produced cardiac hypertrophy. There were three possible mechanisms for the action of the hormone; the investigators judged that they could determine which was the correct one by treating groups of animals with various combinations of the hormone and other compounds and observing which ones altered the action of the hormone. There were ten treatment groups, the active treatments being captopril, hydralazine (two reducing agents), propranolol (a beta adrenoceptor), and a combination of captopril and propranolol, each given with and without T4 (thyroid hormone); in addition there was a group receiving T4 alone and an untreated control group. The treatment combinations actually formed a 5×2 factorial design with unequal sample sizes. However, the investigators analysed the data as two sets (or 'families') of treatments versus specified treatment multiple comparisons:

1. the group receiving T4 alone and each of the 'without T4' treatment groups were compared with the untreated controls; and
2. the various 'with T4' groups were compared with the T4 alone group.

Five cardiovascular variables were measured on each animal to assess the effects of the treatments.

In Table I we show the sample sizes and treatment means for one of the response variables (left ventricular weight/body weight measured in mg/g, denoted by LV/body weight) as given in

Table I. Sample sizes, means, standard deviations and t-statistics for LV/body weight (in mg/g) data from Bedotto *et al.*[18]

Treatment group	Sample size	Sample mean	Sample standard deviation	t-statistics for differences	
				from controls	from T4
Control (untreated)	11	2·21	0·137	—	—
T4	10	2·52	0·175	4·57	—
Captopril	10	2·03	0·203	− 2·75	—
Propranolol	12	2·32	0·162	1·74	—
Hydralazine	10	2·10	0·099	− 1·62	—
Propranolol + Captopril	9	2·04	0·155	− 2·52	—
T4 + Captopril	9	2·49	0·090	—	− 0·36
T4 + Propranolol	12	2.60	0·152	—	1·32
T4 + Hydralazine	10	2·54	0·158	—	0·35
T4 + Propranolol + Captopril	10	2·56	0·167	—	0·55

Bedotto *et al.*,[18] along with the standard deviations for the individual treatment groups which were obtained directly from the investigators.

We first discuss the use of a pooled variance estimate as called for in the description of the proposed procedure. This assumes that the different treatments in the experiment affect only the mean values of response, and have no effect on the variances. If, in fact, the variances are also different in the treatment groups, and if this is ignored and a pooled variance estimate is used, not only will this result in a distortion of the values of the individual t-statistics but also the actual error rate can be quite different from its nominal value. (For a discussion of some studies of the effects of using a pooled variance estimate when the variances are heterogeneous on the error rates of pairwise comparison procedures, see Chapter 7 of Hochberg and Tamhane.[12]) Thus, it is important to check the equal variances assumption. In the present example, we applied Bartlett's test which indicated that the sample variances were not significantly different from each other. Therefore, we used a pooled variance estimate $s^2 = 0{\cdot}02366$ with 93 d.f. to compute the Student t-statistics shown in Table I.

If the equal variances assumption is not justified, the experimental procedure should be examined to determine whether any 'assignable cause' can be found other than the difference in treatments. If this possibility is ruled out, one should try a suitable transformation of the data to see whether the assumption can be met on the transformed scale. Failing this, one alternative would be to apply the Holm[1] procedure, but with t-tests using a pooled variance estimate from only the treatment-control pair under consideration, or approximate t-tests using separate variance estimates. Another alternative would be to apply the approximate form of the present test procedure using separate variance estimates as suggested by Dunnett.[14]

Table II shows the computations of p-values for the first family of comparisons. For two-sided tests, the t-statistics are ordered according to their magnitudes and assigned an index m according to their rank orders. The λ-values needed to compute the m-variate probabilities p'_m depend on the sample sizes and are calculated using the formula for $\lambda_m = 1/\sqrt{(1 + n_0/n_m)}$. The p'_m-values are given by eqn. (5) and the values shown in Table II were computed using the program described in Section 3. Finally, the p-values corresponding to each treatment were determined from the p'-values as shown in eqn. (4). It may be noted that, in this example, only one of the p-values differs from the corresponding p'-value.

Table II. Computation of p-values for comparisons with untreated control

Index m	Ordered $\lvert t_m \rvert$	λ_m	mp''_m	p'_m	p_m
5	4·57	0·6901	0·000	0·000	0·000
4	2·75	0·6901	0·029	0·025	0·025
3	2·52	0·6708	0·040	0·037	0·037
2	1·74	0·7223	0·170	0·151	0·151
1	1·62	0·6901	0·109	0·109	0·151

Table III. Comparison of exact and approximate fixed $\alpha = 0{\cdot}05$ critical values for comparisons with untreated control

Index m	$\bar{\rho}_m$	Upper 5% critical points		
		Exact	$\bar{\rho}$-approx.	Bonferroni approx.
5	0·4797	2·562	2·563	2·630
4	0·4806	2·489	2·489	2·547
3	0·4820	2·391	2·391	2·438
2	0·4984	2·246	2·246	2·278
1	—	1·986	1·986	1·986

Next, we illustrate the application of the fixed-α step-down procedure, using $\alpha = 0{\cdot}05$ as an example. Of course, if the p-values have already been computed as described above, it is only necessary to observe which of them are less than or equal to the specified α, in this case those for treatments 5, 4 and 3 (T4, captopril, and captopril + propranolol, respectively). The alternative method is to determine critical values for each $\lvert t_m \rvert$ for the chosen level $\alpha = 0{\cdot}05$.

Table III shows the exact critical values for each comparison, computed by evaluating the multivariate probabilities, along with the $\bar{\rho}$ and Bonferroni approximations. To determine the $\bar{\rho}$ approximation, we average the ρ's determined by the λ's for each m. For example, for $m = 5$, there are 10 correlation coefficients to be averaged, for $m = 4$, there are 6, and so on. These averages are shown in Table III. Then the approximate critical values for each m were obtained from Table B.3 of Bechhofer and Dunnett[9] using the value of m for the parameter p in that table and interpolating with respect to the value of $\bar{\rho}$ obtained for each m and with respect to $\nu = 93$ d.f. (Interpolation was done linearly in $1/(1-\bar{\rho})$ and $1/\nu$ for greatest accuracy.) To obtain the Bonferroni approximations, the $0{\cdot}05/m$ critical values of univariate Student's t are needed, which may be most easily obtained from Bailey.[16] Whichever method is used to obtain the critical values, it is only necessary to compute them for $m = 5, 4$, etc. until a non-significant result is observed. In this case, we could stop at $m = 2$ but, for illustrative purposes, we show the values obtained for all m.

It may be noted that the $\bar{\rho}$ approximation gives particularly accurate results compared with the exact values in this example because the n's do not differ much, making the ρ_{ij}'s approximately equal. The conclusion from the results shown in Tables II and III is that three of the five treatments in the first family of comparisons (T4, captopril and captopril + propranolol) have demonstrated significant effects on LV/body weight relative to the results obtained for the untreated animals. For the remaining two treatments (propranolol and hydralazine), the effects were non-significant (which does not mean they were necessarily negligible, of course).

Table IV. Computation of p-values and lower 95 per cent simultaneous confidence limits for comparisons with T4

Index m	Ordered $\lvert t_m \rvert$	λ_m	p_m	Lower confidence limit
4	1·32	0·7385	0·49	−0·06
3	0·55	0·7071	0·90	−0·11
2	0·36	0·6882	0·91	−0·18
1	0·35	0·7071	0·91	−0·13

Computations similar to those shown in Tables II and III can be performed for the second family of comparisons as well. In Table IV, we show the p-values obtained for this family. It is clear that none of these treatment groups differs significantly from the T4 alone group for this response variable, that is, H_{0i}: $\mu_i - \mu_0 = 0$ cannot be rejected for $i = 1, \ldots, 4$. It may be useful, however, to determine how close to 0 the various $\mu_i - \mu_0$ are estimated to lie. For this purpose, we consider confidence limits for the $\mu_i - \mu_0$. Since the investigators were particularly interested in looking for differences which would indicate a blocking of the action of T4, which for this variable was to increase it, we show in Table IV the lower 95 per cent simultaneous confidence limits for each $\mu_i - \mu_0$ computed using Dunnett's[2] method. This lower confidence limit is given by

$$\bar{x}_i - \bar{x}_0 - t^{(\alpha)}_{k,\nu,R_k} s \sqrt{(1/n_i + 1/n_0)}$$

where $t^{(\alpha)}_{k,\nu,R_k} = 2{\cdot}19$ for $\alpha = 0{\cdot}05$, $k = 4$, $\nu = 93$ and the R_k matrix is given by the λ's shown in Table IV. From these results we can state, with 95 per cent confidence, that none of the treatments decreased the effect of T4 by more than 0·18 mg/g. It would also be possible to calculate p-values to test for effects in excess of a specified level $-\delta$, the value chosen for the latter being defined as the minimum difference between treatment effects considered to be clinically significant. The hypotheses tested would be formulated as H_{0i}: $\mu_i - \mu_0 \leqslant -\delta$ against upper one-sided alternatives H_{1i}: $\mu_i - \mu_0 > -\delta$. The p-values could be calculated for each comparison as before, small values indicating evidence supporting the conclusion of negligible effect (that is, equivalence of the effects of T4 + drug and T4 alone; see Metzler[19]).

5. DISCUSSION

In this paper we have described a step-down procedure for comparing several treatments with a specified treatment. This problem is a common one in biopharmaceutical and medical studies. The step-down feature of the procedure is not new, as we have pointed out. What is new is the application of the method to unbalanced data, which is made possible by the availability of the computer algorithm described in Section 3. Previous papers by other authors have proposed approximations such as the Bonferroni for use in unbalanced situations. However, with the use of the alternative methods described in the paper, namely use of the computer program or the approximate calculation using tables, exact or virtually exact results are made possible. With the computer program, the method provides p-values, which are more useful than determining whether a particular significance level α is reached.

A limitation of the method is that it is for significance testing only. In applications where the magnitudes of treatment differences are important, either to show how large these might be or to demonstrate that they cannot be larger than some specified amount, simultaneous confidence

intervals will be more appropriate. For simultaneous confidence interval estimates of differences between treatments and a specified treatment, see Dunnett[2] and Hochberg and Tamhane[12] (pp. 140–141), and note that the same methods we have described here can be applied in that context for unbalanced data.

ACKNOWLEDGEMENTS

C. W. Dunnett's research was supported by a grant from the Natural Sciences and Engineering Research Council of Canada at McMaster University. The authors are indebted to Dr. Richard Gay for providing additional data from the study reported in Bedotto *et al.*,[18] which enabled us to calculate the statistics presented in Section 4 with greater precision.

REFERENCES

1. Holm, S. 'A simple sequentially rejective multiple test procedure,' *Scandinavian Journal of Statistics*, **6**, 65–70 (1979).
2. Dunnett, C. W. 'A multiple comparison procedure for comparing several treatments with a control', *Journal of the American Statistical Association*, **50**, 1096–1121 (1955).
3. Roessler, E. B. 'Testing the significance of observations compared with a control', *Proceedings of the American Society for Horticultural Science*, **47**, 249–251 (1946).
4. Slepian, D. 'The one-sided barrier problem for Gaussian noise', *Bell System Technical Journal*, **41**, 463–501 (1962).
5. Šidák, Z. 'Rectangular confidence regions for the means of multivariate normal distributions', *Journal of the American Statistical Association*, **62**, 626–633 (1967).
6. Miller, R. G., Jr. *Simultaneous Statistical Inference*, McGraw Hill, New York, 1966.
7. Naik, U. D. 'Some selection rules for comparing p processes with a standard', *Communications in Statistics – A. Theory and Methods*, **4**, 519–535 (1975).
8. Marcus, R., Peritz, E. and Gabriel, K. R. 'On closed testing procedures with special reference to ordered analysis of variance', *Biometrika*, **63**, 655–660 (1976).
9. Bechhofer, R. E. and Dunnett, C. W. 'Percentage points of multivariate Student t distributions', *Selected Tables in Mathematical Statistics*, **11**, Providence: American Mathematical Society (1988).
10. Dunnett, C. W. 'Multivariate normal probability integrals with product correlation structure, Algorithm AS 251', *Applied Statistics*, **38**, 564–579 (1989).
11. Westfall, P. H. and Young, S. 'P value adjustments for multiple tests in multivariate binomial models', *Journal of the American Statistical Association*, **84**, 780–786 (1989).
12. Hochberg, Y. and Tamhane, A. C. *Multiple Comparison Procedures*, Wiley, New York, 1987.
13. Curnow, R. N. and Dunnett, C. W. 'The numerical evaluation of certain multivariate normal integrals', *Annals of Mathematical Statistics*, **33**, 571–579 (1962).
14. Dunnett, C. W. 'Multiple comparisons between several treatments and a specified treatment', Invited talk at the Spring ENAR Meeting, Raleigh, NC, 1985.
15. Dutt, J. E., Mattes, K. D., Soms, A. P. and Tao, L. C. 'An approximation to the trivariate t with a comparison to the exact values', *Biometrics*, **41**, 153–169 (1976).
16. Bailey, B. J. R. 'Tables of the Bonferroni t statistic', *Journal of the American Statistical Association*, **72**, 469–478 (1977).
17. Holland, B. S. and Copenhaver, M. D. 'An improved sequentially rejective Bonferroni test procedure', *Biometrics*, **43**, 417–424 (1987).
18. Bedotto, J. E., Gay, R. G., Graham, S. D., Morkin, E. and Goldman, S. 'Cardiac hypertrophy induced by thyroid hormone is independent of loading conditions and beta adrenoceptor', *Journal of Pharmacology and Experimental Therapeutics*, **248**, 632–636 (1989).
19. Metzler, C. M. 'Sample sizes for bioequivalence studies', *Statistics in Medicine*, **10**, 961–970 (1991).