The Canadian Journal of Statistics
Vol. 26, No. 2, 1998, Pages 353–363
La Revue Canadienne de Statistique

A generalized step-up-down multiple test procedure

Ajit C. TAMHANE, Wei LIU and Charles W. DUNNETT*

Northwestern University, University of Southampton and *McMaster University*

Key words and phrases: Biometric applications, multiple comparisons, stepwise test procedures, familywise error rate, power.
AMS 1991 subject classifications: 62F03, 92C50.

ABSTRACT

A generalization of step-up and step-down multiple test procedures is proposed. This step-up-down procedure is useful when the objective is to reject a specified minimum number, q, out of a family of k hypotheses. If this basic objective is met at the first step, then it proceeds in a step-down manner to see if more than q hypotheses can be rejected. Otherwise it proceeds in a step-up manner to see if some number less than q hypotheses can be rejected. The usual step-down procedure is the special case where $q = 1$, and the usual step-up procedure is the special case where $q = k$. Analytical and numerical comparisons between the powers of the step-up-down procedures with different choices of q are made to see how these powers depend on the actual number of false hypotheses. Examples of application include comparing the efficacy of a treatment to a control for multiple endpoints and testing the sensitivity of a clinical trial for comparing the efficacy of a new treatment with a set of standard treatments.

RÉSUMÉ

Une généralisation des procédures de test multiples step-up et step-down est proposée. Cette procédure step-up-down est utile lorsque l'objectif est de rejeter un nombre minimum spécifié q d'une famille de k hypothèses. Si l'objectif de base est atteint à la première étape, alors elle procède dans une manière descendante pour voir si plus que q hypothèses peuvent être rejetées. Sinon, elle procède d'une manière ascendante pour voir si un nombre inférieur à q d'hypothèses peut être rejeté. La procédure step-down habituelle est le cas spécial où $q = 1$, et la procédure step-up habituelle est le cas spécial où $q = k$. Des comparaisons analytiques et numériques entre les puissances des procédures step-up-down avec différents choix de q sont proposées afin de montrer comment ces puissances dépendent de nombre d'hypothèses fausses. Des exemples d'application incluent la comparaison de l'efficacité d'un traitement afin de contrôler les point finaux multiples et de tester la sensibilité d'un essai clinique comparant l'efficacité l'efficacité d'un nouveau traitement avec un ensemble de traitements standard.

1. INTRODUCTION

In this paper, we consider the problem of simultaneously testing a family of $k \geq 2$ hypotheses, $\mathcal{H}_1, \mathcal{H}_2, \ldots, \mathcal{H}_k$, based on test statistics $t_1, t_2, \ldots, t_k$, respectively. The familywise error rate is defined as the probability of rejecting at least one true $\mathcal{H}_i$. We require that any multiple test procedure for this problem control the familywise error rate at a designated level α, irrespective of which and how many of the $\mathcal{H}_i$ are true. A multiple test procedure satisfying this condition will be referred to as an α-level multiple

*Supported by a research grant from the Natural Sciences and Engineering Research Council of Canada.

test procedure. See Hochberg and Tamhane (1987) for a rationale behind control of the familywise error rate and for a general introduction to multiple comparison problems.

There are two types of multiple test procedures: single-step and stepwise. In a single-step procedure the decision on any $\mathcal{H}_i$ does not depend upon the decision on any other $\mathcal{H}_j$; therefore the hypotheses can be tested without reference to one another. A single-step testing procedure typically arises when a simultaneous confidence-interval procedure is used to make the hypothesis tests. On the other hand, in a stepwise procedure the hypotheses are tested step by step in some order. In the present paper, this order is determined by the ordered values of the test statistics, $t_{(1)} \leq t_{(2)} \leq \cdots \leq t_{(k)}$, where a larger value of a test statistic is assumed to be more significant. The decisions on the earlier hypotheses in the order tested affect the decisions on the hypotheses tested later. Stepwise procedures are more powerful than single-step procedures when the detection of more than one false hypothesis is of interest.

There are two types of stepwise procedures: step-down and step-up. To explain these procedures, we introduce the notation $\mathcal{H}_{(i)}$ for the hypothesis associated with the ith ordered test statistic, $t_{(i)}$. In a step-down procedure testing begins with $\mathcal{H}_{(k)}$, i.e., the hypothesis associated with the most significant test statistic, $t_{(k)} = t_{\max}$. If $H_{(k)}$ is rejected, testing continues in the order $\mathcal{H}_{(k-1)}, \mathcal{H}_{(k-2)}, \ldots$, as long as rejection occurs at each step. Testing stops either when there are no more hypotheses to test or when a hypothesis is not rejected ("accepted"), at which point the remaining hypotheses are accepted by implication without actually testing them. In a step-up procedure testing begins with $\mathcal{H}_{(1)}$, i.e., the hypothesis associated with the least significant test statistic, $t_{(1)} = t_{\min}$. If $\mathcal{H}_{(1)}$ is accepted, testing continues in the order $\mathcal{H}_{(2)}, \mathcal{H}_{(3)}, \ldots$, as long as acceptance occurs at each step. Testing stops either when there are no more hypotheses to test or when a hypothesis is rejected, at which point the remaining hypotheses are rejected by implication without actually testing them.

The test in the first step of a step-down procedure answers the question "Can at least one hypothesis be rejected?" by using a test based on $t_{\max}$. If the answer to this question is affirmative, then the procedure proceeds testing in a step-down manner to provide a further resolution of this question. The test in the first step of a step-up procedure answers the question "Can all hypotheses be rejected?" by using a test based on $t_{\min}$ [the "min" test of Laska and Meisner (1989)]. If the answer to this question is negative, then the procedure proceeds testing in a step-up manner to provide a further resolution of this question.

In some applications the investigator wants to answer the question "Can at least q hypotheses be rejected?" where q is a specified integer between 1 and k. For example, a new treatment may be preferred to a placebo if it shows efficacy on at least q out of k endpoints; another example is given in Section 6. A test to answer this question can be based on the statistic $t_{(r)}$ where $r = k - q + 1$ (reject $\mathcal{H}_{(r)}, \mathcal{H}_{(r+1)}, \ldots, \mathcal{H}_{(k)}$ if $t_{(r)} > c_r$). A stepwise extension of this test proceeds in a step-down or step-up manner depending on whether the answer to the question is affirmative or not. We call the resulting stepwise procedure a step-up-down procedure and denote it by SUDP(r). The step-down and step-up procedures are special cases of SUDP(r) for $q = 1$, $r = k$ and $q = k$, $r = 1$, respectively.

The purpose of the present paper is to develop this generalized procedure in the context of the normal-theory linear model for a balanced design. A study of the power properties of this procedure shows that if m is the actual number of true hypotheses and $q = k - m$ is the actual number of false hypotheses, then the most powerful SUDP(r) is obtained when $r = m + 1 = k - q + 1$. Thus another application of SUDP(r) is when one has some prior idea about the likely number of true/false hypotheses.

The plan of the paper is as follows. Section 2 gives the distributional setup for balanced designs. For this setup the estimates of the test parameters have equal variances and equal correlation coefficients, which simplifies the calculation of the critical constants of SUDP(r). Section 3 states the procedure. Section 4 shows how to obtain its critical constants. A table of selected constants is provided. Section 5 gives results, both analytical and numerical, on the power properties of SUDP(r). Section 6 discusses an application dealing with demonstrating the sensitivity of a clinical trial for a new generic drug. Finally, Section 7 gives concluding remarks.

2. PRELIMINARIES

Consider the following one-sided multiple hypotheses testing problem on k unknown parameters, $\theta_1, \theta_2, \ldots, \theta_k$:

$$\mathcal{H}_i : \theta_i = 0 \text{ vs. } A_i : \theta_i > 0 \qquad (1 \le i \le k). \tag{2.1}$$

We assume the standard-normal-theory setup where the θ_i are estimable linear parametric functions (typically contrasts among the treatment means). Let $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_k$ be their least-squares estimates. Suppose that the design is balanced in the sense that the $\hat{\theta}_i$ have equal variances and equal correlation coefficients. Specifically, we assume that the $\hat{\theta}_i$ have a joint k-variate normal distribution with

$$\mathcal{E}\,\hat{\theta}_i = \theta_i, \qquad \mathcal{V}ar\ \hat{\theta}_i = \tau^2\sigma^2 \quad \text{and} \quad Corr\,(\hat{\theta}_i, \hat{\theta}_j) = \rho \quad \text{for all } i \ne j; \tag{2.2}$$

here τ^2 and ρ are known design-dependent constants, and σ^2 is an unknown experimental error variance. Let s^2 be an estimate of σ^2 based on ν degrees of freedom (d.f.), so that the corresponding random variable (r.v.) S^2 is distributed as $\sigma^2\chi^2_\nu/\nu$ independently of the $\hat{\theta}_i$.

Three examples of this setup are: (1) comparisons of treatments with a control in a one-way layout (Dunnett 1955, 1997) with an equal number, n, of observations on each treatment and possibly a different number, n_0, of observations on the control; (2) orthogonal contrasts among the cell means corresponding to main effects and interactions in a two-level factorial experiment with equireplicated cells; and (3) a BTIB design (Bechhofer and Tamhane 1981) for comparing treatments with a control using incomplete blocks.

The test statistics used to test the $\mathcal{H}_i$ are

$$t_i = \frac{\hat{\theta}_i}{\text{SE}(\hat{\theta}_i)} = \frac{\hat{\theta}_i}{s\tau} \qquad (1 \le i \le k). \tag{2.3}$$

The r.v.'s T_i corresponding to the observed statistics t_i have a k-variate t-distribution with common correlation ρ and d.f. ν; the subset of the T_i corresponding to the true $\mathcal{H}_i$ has a central t-distribution, and the complementary subset has a noncentral t-distribution. Denote by $t^{(\alpha)}_{k,\nu,\rho}$, the upper α equicoordinate critical point of a central k-variate t-distribution with common correlation ρ and d.f. ν. These critical points will be needed in (4.2) below. Comprehensive tables of $t^{(\alpha)}_{k,\nu,\rho}$ are given in Bechhofer and Dunnett (1988), or they can be computed using the algorithm of Dunnett (1989).

3. A GENERALIZED STEP-UP-DOWN PROCEDURE

For fixed integer r ($1 \le r \le k$), the step-up-down procedure SUDP(r) begins by testing the hypothesis $\mathcal{H}_{(r)}$ that corresponds to the rth test statistic in ascending order of

significance. If $\mathcal{H}_{(r)}$ is rejected, then $\mathcal{H}_{(r+1)},\ldots,\mathcal{H}_{(k)}$ are also rejected and testing continues by testing $\mathcal{H}_{(r-1)},\mathcal{H}_{(r-2)},\ldots$, in the usual step-down manner until a hypothesis is accepted. If $\mathcal{H}_{(r)}$ is accepted, then $\mathcal{H}_{(1)},\ldots,\mathcal{H}_{(r-1)}$ are also accepted and testing continues by testing $\mathcal{H}_{(r+1)},\mathcal{H}_{(r+2)},\ldots$, in the usual step-up manner until a hypothesis is rejected.

The steps in SUDP(r) are as follows:

Step 0. Order the test statistics t_i: $t_{(1)} \leq t_{(2)} \leq \cdots \leq t_{(k)}$. Let $\mathcal{H}_{(1)},\mathcal{H}_{(2)},\ldots,\mathcal{H}_{(k)}$ be the corresponding hypotheses. Choose critical constants $c_1 \leq c_2 \leq \cdots \leq c_k$ as indicated in (4.2) below.

Step 1(a). If $t_{(r)} \leq c_r$ then accept $\mathcal{H}_{(1)},\mathcal{H}_{(2)},\ldots,\mathcal{H}_{(r)}$ and go to general step (a).

Step 1(b). If $t_{(r)} > c_r$ then reject $\mathcal{H}_{(r)},\mathcal{H}_{(r+1)},\ldots,\mathcal{H}_{(k)}$ and go to general step (b).

General step (a). Let $\mathcal{H}_{(i)}$ denote the last accepted hypothesis [at Step 1(a), $i = r$]. If $i = k$ then stop testing; otherwise test $\mathcal{H}_{(i+1)}$. If $t_{(i+1)} > c_{i+1}$ then reject $\mathcal{H}_{(i+1)},\mathcal{H}_{(i+2)},\ldots,\mathcal{H}_{(k)}$ and stop testing. If $t_{(i+1)} \leq c_{i+1}$ then accept $\mathcal{H}_{(i+1)}$. Set i to $i+1$ and return to the beginning of this step.

General step (b). Let $\mathcal{H}_{(i)}$ denote the last rejected hypothesis [at step 1(b), $i = r$]. If $i = 1$ then stop testing; otherwise test $\mathcal{H}_{(i-1)}$. If $t_{(i-1)} \leq c_{i-1}$ then accept $\mathcal{H}_{(i-1)},\mathcal{H}_{(i-2)},\ldots,\mathcal{H}_{(1)}$ and stop testing. If $t_{(i-1)} > c_{i-1}$ then reject $\mathcal{H}_{(i-1)}$. Set i to $i-1$ and return to the beginning of this step.

For $r = 1$, SUDP(r) reduces to the step-up procedure of Dunnett and Tamhane (1992a), while for $r = k$, SUDP(r) reduces to the step-down procedure given variously by Miller (1966), Naik (1975), Marcus *et al.* (1976) and Dunnett and Tamhane (1991).

4. DETERMINATION OF CRITICAL CONSTANTS

We determine the critical constants of SUDP(r) so that the requirement that the family-wise error rate be $\leq \alpha$ is met. Let $\boldsymbol{\theta}_m$ be any parameter configuration $\boldsymbol{\theta} = (\theta_1,\ldots,\theta_k)$ such that $\theta_i = 0$ for $i = 1,\ldots,m$ and $\theta_i > 0$ for $i = m+1,\ldots,k$; thus $\mathcal{H}_1,\ldots,\mathcal{H}_m$ are true and $\mathcal{H}_{m+1},\ldots,\mathcal{H}_k$ are false ($m = 1,2,\ldots,k$).

THEOREM 4.1. *The critical constants $c_1 \leq c_2 \leq \cdots \leq c_m$ required to satisfy*

$$P_{\boldsymbol{\theta}_m}(\text{accept } \mathcal{H}_1,\ldots,\mathcal{H}_m) \geq 1-\alpha \qquad \text{for } m = 1,\ldots,k \tag{4.1}$$

are obtained by solving the equation

$$P\{(T_1,T_2,\ldots,T_m) \leq (\underbrace{c_r,\ldots,c_r}_{r},c_{r+1},\ldots,c_m)\} = 1-\alpha \quad \text{for} \quad m = r+1,\ldots,k, \tag{4.2}$$

where $c_r = t^{(\alpha)}_{r,\nu,\rho}$ and the notation $(T_1,T_2,\ldots,T_m) \leq (c_1,c_2,\ldots,c_m)$ means that the smallest of the T_i is $\leq c_1$, the next smallest is $\leq c_2$, and so on.

Proof. Available from the authors. □

Equation (4.2) can be solved recursively starting with $m = r+1$, then $m = r+2$ and so on. Note that $c_1 = t^{(\alpha)}_{\nu}$, the upper α critical point of Student's t with ν d.f., for all values of r. For $r = k$, we have $c_m = t^{(\alpha)}_{m,\nu,\rho}$ for $m = 1,2,\ldots,k$, which are the critical constants used by the step-down procedure, and for $r = 1$, the c_m are determined recursively from the equation

$$P\{(T_1,T_2,\ldots,T_m) \leq (c_1,c_2,\ldots,c_m)\} = 1-\alpha \qquad \text{for} \quad m = 1,2,\ldots,k,$$

which are the critical constants used by the step-up procedure.

We now give an expression for the probability in (4.2) that can be evaluated efficiently on a computer. Let Z_i for $i = 0, 1, \ldots, k$ be independent $\mathsf{N}(0, 1)$ r.v.'s, and let U be a $\sqrt{\chi^2_\nu/\nu}$ r.v. independent of the Z_i. Then under $\boldsymbol{\theta} = \boldsymbol{\theta}_k = \mathbf{0}$ we can express $T_i = (\sqrt{1-\rho}\, Z_i - \sqrt{\rho}\, Z_0)/U$ $(1 \le i \le m)$ and write the desired probability as

$$\int_0^\infty \int_{-\infty}^\infty P\{(Z_1, \ldots, Z_m) \le (\underbrace{d_r, \ldots, d_r}_{r}, d_{r+1}, \ldots, d_m)\}\phi(z_0) f_\nu(u)\, dz_0\, du, \tag{4.3}$$

where $\phi(\cdot)$ is the standard normal density function, $f_\nu(\cdot)$ is the density function of U, and $d_i = (c_i u + \sqrt{\rho} z_0)/\sqrt{1-\rho}$. The probability

$$P\{(Z_1, \ldots, Z_m) \le (\underbrace{d_r, \ldots, d_r}_{r}, d_{r+1}, \ldots, d_m)\}$$

can be evaluated recursively using Lemma 3.1 of Dunnett and Tamhane (1992a), and then a two-dimensional numerical integration can be employed to evaluate (4.3).

The values of c_m for $k = 1(1)6$, $r = 1(1)k$, $\alpha = 0.05$, $\rho = 0$, 0.25, 0.50 and $\nu = 10$, ∞ are given in Table 1 for the one-sided tests discussed here. More extensive tables and Fortran programs for computing the c_m are available from the authors upon request.

To emphasize the dependence of c_m on r, from now on we will use the notation $c_m(r)$, which are the critical constants used by SUDP(r). It follows that

$$c_m(m) = c_m(m+1) = \cdots = c_m(k) = t^{(\alpha)}_{m,\nu,\rho}. \tag{4.4}$$

The following two relations are also found to hold empirically for all the computations that we have done:

$$c_1(r) \le c_2(r) \le \cdots \le c_k(r) \tag{4.5}$$

and

$$c_m(1) \ge c_m(2) \ge \cdots \ge c_m(m). \tag{4.6}$$

5. POWER OF SUDP(r)

We consider two definitions of power:

$$\pi_1 = P\{\text{reject all false } \mathcal{H}_i\text{'s and accept all true } \mathcal{H}_i\text{'s}\}$$

and

$$\pi_2 = P\{\text{reject all false } \mathcal{H}_i\text{'s}\}.$$

Clearly, $\pi_1 \le \pi_2$. Note that π_1 gives the probability of correct inferences, while π_2 is a more commonly used definition of power. The former definition turns out to be analytically more tractable.

Let $\delta_i = \theta_i/\sigma\tau$ for $1 \le i \le k$, and let $\boldsymbol{\delta}_m = \boldsymbol{\theta}_m/\sigma\tau$ denote the vector of the δ_i's whose first m components are 0 and last $k-m$ components are positive; i.e., the first m hypotheses are true and the remaining $k-m$ are false. The powers π_1 and π_2 of SUDP(r) depend on the θ_i's only through the δ_i's, and will be denoted by $\pi_1(\boldsymbol{\delta}_m|r)$ and $\pi_2(\boldsymbol{\delta}_m|r)$, respectively.

TABLE 1: Critical constants for SUDP(r) for one-sided tests ($\alpha = 0.05$).

ρ	d.f.	r	c_1	c_2	c_3	c_4	c_5	c_6
0.00	10	1	1.812	2.220	2.442	2.600	2.722	2.821
		2	1.812	2.211	2.442	2.600	2.722	2.821
		3	1.812	2.211	2.439	2.600	2.722	2.821
		4	1.812	2.211	2.439	2.598	2.722	2.821
		5	1.812	2.211	2.439	2.598	2.720	2.821
		6	1.812	2.211	2.439	2.598	2.720	2.820
	∞	1	1.645	1.960	2.123	2.235	2.319	2.386
		2	1.645	1.954	2.123	2.235	2.319	2.386
		3	1.645	1.954	2.121	2.235	2.319	2.386
		4	1.645	1.954	2.121	2.234	2.319	2.386
		5	1.645	1.954	2.121	2.234	2.319	2.386
		6	1.645	1.954	2.121	2.234	2.319	2.386
0.25	10	1	1.812	2.205	2.410	2.554	2.665	2.755
		2	1.812	2.189	2.408	2.553	2.664	2.755
		3	1.812	2.189	2.402	2.553	2.664	2.754
		4	1.812	2.189	2.402	2.549	2.664	2.754
		5	1.812	2.189	2.402	2.549	2.662	2.754
		6	1.812	2.189	2.402	2.549	2.662	2.753
	∞	1	1.645	1.953	2.108	2.214	2.295	2.359
		2	1.645	1.942	2.107	2.214	2.295	2.359
		3	1.645	1.942	2.103	2.214	2.295	2.359
		4	1.645	1.942	2.103	2.212	2.295	2.359
		5	1.645	1.942	2.103	2.212	2.293	2.359
		6	1.645	1.942	2.103	2.212	2.293	2.358
0.50	10	1	1.812	2.174	2.350	2.473	2.567	2.643
		2	1.812	2.151	2.347	2.472	2.566	2.642
		3	1.812	2.151	2.337	2.471	2.566	2.642
		4	1.812	2.151	2.337	2.466	2.565	2.642
		5	1.812	2.151	2.337	2.466	2.562	2.642
		6	1.812	2.151	2.337	2.466	2.562	2.640
	∞	1	1.645	1.933	2.071	2.165	2.237	2.294
		2	1.645	1.916	2.068	2.164	2.236	2.294
		3	1.645	1.916	2.062	2.164	2.236	2.294
		4	1.645	1.916	2.062	2.160	2.236	2.294
		5	1.645	1.916	2.062	2.160	2.234	2.294
		6	1.645	1.916	2.062	2.160	2.234	2.292

THEOREM 5.1. *If the T_i's are independent (i.e., if $\rho = 0$ and $\nu = \infty$) and if* (4.5) *and* (4.6) *hold, then for $m = 0, 1, \ldots, k-1$,*

$$\pi_1(\boldsymbol{\delta}_m|1) \leq \cdots \leq \pi_1(\boldsymbol{\delta}_m|m) \leq \pi_1(\boldsymbol{\delta}_m|m+1) \geq \pi_1(\boldsymbol{\delta}_m|m+2) \geq \cdots \geq \pi_1(\boldsymbol{\delta}_m|k). \quad (5.1)$$

Proof. Available from the authors. □

This result shows that when the T_i are independent, if m hypotheses are true and $q = k - m$ are false, then $r = m + 1 = k - q + 1$ yields the most powerful SUDP(r). For the case of dependent T_i's we were able to establish only the following partial result, although numerical evaluations of π_1 (see Table 2) suggest that Theorem 5.1 holds in this case, too.

TABLE 2: Power, $\pi_1(\boldsymbol{\delta}_m|r)$, of SUDP($r$) for $k = 5$, $\nu = \infty$, $\alpha = 0.05$ and $\delta = 3$.

		$\pi_1(\boldsymbol{\delta}_m\|r)$				
ρ	r	$m = 0$	1	2	3	4
0.00	1	0.6319[a]	0.4979	0.5044	0.5748	0.7145
	2	0.6080	0.5011[a]	0.5044	0.5748	0.7145
	3	0.6053	0.4943	0.5054[a]	0.5748	0.7145
	4	0.6051	0.4938	0.5031	0.5752[a]	0.7145
	5	0.6051	0.4938	0.5030	0.5746	0.7145[a]
0.25	1	0.6824[a]	0.5456	0.5429	0.5951	0.7140
	2	0.6608	0.5506[a]	0.5434	0.5951	0.7140
	3	0.6578	0.5444	0.5452[a]	0.5953	0.7140
	4	0.6575	0.5438	0.5429	0.5961[a]	0.7141
	5	0.6575	0.5438	0.5428	0.5954	0.7145[a]
0.50	1	0.7364[a]	0.6072	0.5999	0.6357	0.7279
	2	0.7182	0.6143[a]	0.6012	0.6360	0.7282
	3	0.7153	0.6089	0.6037[a]	0.6361	0.7282
	4	0.7149	0.6083	0.6017	0.6376[a]	0.7282
	5	0.7149	0.6083	0.6016	0.6369	0.7288[a]

[a] Highest power among all SUDP(r) for the given configuration $\boldsymbol{\delta}_m$.

THEOREM 5.2. *If* (4.5) *and* (4.6) *hold, then for* $m = 0, 1, \cdots, k-1$

$$\pi_1(\boldsymbol{\delta}_m|m) \leq \pi_1(\boldsymbol{\delta}_m|m+1) \geq \pi_1(\boldsymbol{\delta}_m|m+2) \geq \cdots \geq \pi_1(\boldsymbol{\delta}_m|k). \tag{5.2}$$

Proof. Available from the authors. □

For $m = 0$, (5.2) yields

$$\pi_1(\boldsymbol{\delta}_0|1) \geq \pi_1(\boldsymbol{\delta}_0|2) \geq \cdots \geq \pi_1(\boldsymbol{\delta}_0|k). \tag{5.3}$$

Thus when all hypotheses are false, the step-up procedure has the highest power, while the step-down procedure has the lowest power, using π_1 as the definition of the power.

The method of proof of Theorem 5.2 fails for showing that

$$\pi_1(\boldsymbol{\delta}_m|r) \leq \pi_1(\boldsymbol{\delta}_m|r+1) \tag{5.4}$$

for $1 \leq r \leq m-1$ when the T_i's are dependent.

We now turn to the analysis of π_2. First note that $\pi_2(\boldsymbol{\delta}_0|r) = \pi_1(\boldsymbol{\delta}_0|r)$. Therefore, analogous to (5.3) we have

$$\pi_2(\boldsymbol{\delta}_0|1) \geq \pi_2(\boldsymbol{\delta}_0|2) \geq \cdots \geq \pi_2(\boldsymbol{\delta}_0|k). \tag{5.5}$$

We next derive an expression for π_2 that can be used for its numerical evaluation. Write

$$\pi_2(\boldsymbol{\delta}_m|r) = \sum_{j=0}^{m} \binom{m}{j} P\{\text{accept } \mathcal{H}_1, \ldots, \mathcal{H}_j \text{ and reject } \mathcal{H}_{j+1}, \ldots, \mathcal{H}_k\}$$

$$= \begin{cases} \displaystyle\sum_{j=0}^{r-1} \binom{m}{j} P_{j1} + \sum_{j=r}^{m} \binom{m}{j} P_{j2} & \text{if } r \leq m, \\ \displaystyle\sum_{j=0}^{m} \binom{m}{j} P_{j1} & \text{if } r > m, \end{cases}$$

TABLE 3: Power, $\pi_2(\boldsymbol{\delta}_m|r)$, of SUDP($r$) for $k = 5$, $\nu = \infty$, $\alpha = 0.05$ and $\delta = 3$.

		$\pi_2(\boldsymbol{\delta}_m\|r)$				
ρ	r	$m = 0$	1	2	3	4
0.00	1	0.6319[a]	0.5325[a]	0.5356[a]	0.6077[a]	0.7534[a]
	2	0.6080	0.5313	0.5355	0.6077	0.7534
	3	0.6053	0.5240	0.5349	0.6077	0.7534
	4	0.6051	0.5235	0.5324	0.6075	0.7534
	5	0.6051	0.5235	0.5324	0.6068	0.7531
0.25	1	0.6824[a]	0.5907	0.5866	0.6393	0.7604
	2	0.6608	0.5934[a]	0.5870	0.6394	0.7604
	3	0.6578	0.5870	0.5880[a]	0.6395	0.7604
	4	0.6575	0.5864	0.5857	0.6400[a]	0.7604
	5	0.6575	0.5864	0.5856	0.6393	0.7607[a]
0.50	1	0.7364[a]	0.6565	0.6491	0.6849	0.7774
	2	0.7182	0.6631[a]	0.6504	0.6853	0.7777
	3	0.7153	0.6576	0.6526[a]	0.6853	0.7777
	4	0.7149	0.6571	0.6507	0.6867[a]	0.7777
	5	0.7149	0.6570	0.6505	0.6861	0.7783[a]

[a] Highest power among all SUDP(r) for the given configuration $\boldsymbol{\delta}_m$.

where

$$P_{j1} = P\left\{\max_{1\le i\le j} T_i \le c_j;\ (T_{j+1},\ldots,T_k) > (c_{j+1},\ldots,c_{r-1},\underbrace{c_r,\ldots,c_r}_{k-r+1})\right\} \quad \text{for } j < r$$

and

$$P_{j2} = P\left\{(T_1,\ldots,T_j) \le (\underbrace{c_r,\ldots,c_r}_{r},c_{r+1},\ldots,c_j);\ \min_{j+1\le i\le k} T_i > c_{j+1}\right\} \quad \text{for } j \ge r.$$

It appears difficult to establish an analytical result analogous to Theorem 5.1 for π_2. However, the above expressions can be evaluated numerically by expressing $T_i = (\sqrt{1-\rho}Z_i - \sqrt{\rho}Z_0)/U$ as in (4.3).

We have calculated π_1 and π_2 for configurations $\boldsymbol{\delta}_m$ where $\delta_1 = \ldots = \delta_m = 0$ and $\delta_{m+1} = \cdots = \delta_k = \delta > 0$ for $m = 0, 1, \ldots, k-1$. Table 2 gives values of π_1 and Table 3 gives values of π_2 for $k = 5$, $\rho = 0$, 0.25, 0.50, $\alpha = 0.05$, $\nu = \infty$ and $\delta = 3$. From Table 2 we see that, using π_1 as the definition of power, SUDP(r) with $r = m+1$ is the most powerful procedure at all $\boldsymbol{\delta}_m$ for all three values of ρ. On the other hand, from Table 3 we see that, using π_2 as the definition of power, SUDP(1) is the most powerful procedure at all $\boldsymbol{\delta}_m$ for $\rho = 0$, while SUDP($m+1$) is the most powerful procedure at all $\boldsymbol{\delta}_m$ for $\rho = 0.25$ and 0.5. For small values of ρ (e.g., $\rho = 0.1$) the most powerful choice of r is found to be between 1 and $m+1$.

When $m = 0$, i.e., when all hypotheses are false, the most powerful procedure (using either definition of power) is the step-up procedure SUDP(1), which has a moderate power gain over SUDP(r) with $r > 1$. For $m \ge 1$, however, the power gain of the most powerful procedure SUDP(r) over SUDP(1) is negligible, as can be readily seen. These findings are in agreement with those in Dunnett and Tamhane (1993), where the step-up [SUDP(1)] and step-down [SUDP(k)] procedures were compared.

6. AN EXAMPLE

As mentioned in the introduction, in some applications a trial is regarded as a "success" when a specified minimum number, q, of hypotheses are rejected. For example, consider the application described in Dunnett and Tamhane (1992b) which involves comparisons of a test drug with both a placebo and $k \geq 2$ known active controls for efficacy. The primary purpose of the trial is to demonstrate the efficacy of the test drug with respect to the placebo. If the test drug is shown to be effective, then it is also of interest to compare its efficacy with each of the known actives.

However, before proceeding with these comparisons, it is necessary to make preliminary comparisons between the known actives and the placebo to establish the sensitivity of the trial (i.e., the ability of the trial to detect differences between the known actives and the placebo). If the sensitivity of the trial cannot be established, the trial may be judged a failure and subsequent comparisons may be abandoned.

Label the placebo as 0, the known active controls as $1, 2, \ldots, k$, and the test drug as $k+1$. Let $\mu_0, \mu_1, \ldots, \mu_k, \mu_{k+1}$ be the respective mean responses of the $k+2$ treatments. The trial is defined to be sensitive if it detects at least a predetermined number q $(1 \leq q \leq k)$ of the known active drugs to be different from the placebo. Thus we want to test

$$\mathcal{H} : \mu_i - \mu_0 = 0 \text{ for at least } k - q + 1 \text{ values of } i = 1, 2, \ldots, k$$

versus

$$A : \mu_i - \mu_0 > 0 \text{ for at least } q \text{ values of } i = 1, 2, \ldots, k.$$

Provided that $\mathcal{H}$ is rejected, the test drug is next compared with the placebo to show that $\mu_{k+1} - \mu_0 > 0$. If this test is satisfactory, the test drug is finally compared with each of the known active drugs (possibly after deleting any active drugs that failed to show an effect compared to the placebo in the preliminary tests) to determine which differences $\mu_{k+1} - \mu_i$ can be shown to be positive.

Consider the problem of establishing sensitivity by testing $\mathcal{H}$ versus A. In Dunnett and Tamhane (1992b) a single-step test was proposed that rejects $\mathcal{H}$ if $t_{(k-q+1)} > c_{k-q+1}$, where $c_{k-q+1} = t^{(\alpha)}_{k-q+1,\nu,\rho}$. Now SUDP($r$) with $r = k-q+1$ can be viewed as a stepwise extension of this single-step test. Instead of testing a single hypothesis $\mathcal{H}$ versus A, SUDP(r) tests k hypotheses $\mathcal{H}_i$ versus A_i given in (2.1) with $\theta_i = \mu_i - \mu_0$. The first step of SUDP(r) is the single-step test stated above. If $t_{(k-q+1)} > c_{k-q+1}$, then q hypotheses (namely, $\mathcal{H}_{(k-q+1)}, \ldots, \mathcal{H}_{(k)}$) are rejected and the required sensitivity of the trial is established. SUDP(r) does further step-down testing to determine whether any additional known actives can be shown to be effective compared with the placebo. On the other hand, if $t_{(k-q+1)} \leq c_{k-q+1}$, then the hypotheses $\mathcal{H}_{(1)}, \ldots, \mathcal{H}_{(k-q+1)}$ are accepted, so that sensitivity of the trial is not established. However, there may be explanatory reasons for this (e.g., reduced sample sizes due to dropouts, noncompliance, etc.). SUDP(r) does further step-up testing to see whether any of the remaining hypotheses can be rejected.

In the final stage, SUDP(r) can again be used with r based on a number q' of active standards that the test drug should be superior to in order to justify the introduction of the test drug into the market.

7. CONCLUDING REMARKS

The main use of SUDP(r) is in those applications where it is desired to show that at least q out of k null hypotheses are false in which case we choose $r = k - q + 1$. The fact that the most powerful SUDP(r) can be found if the number of true hypotheses, m,

is known *a priori* [in many cases the most powerful SUDP(r) is obtained by choosing $r = m + 1$] is mainly of theoretical interest, for two reasons. First, such knowledge is rarely available. Second, the choice $r = 1$ achieves nearly the highest power in all cases, and therefore the step-up procedure can always be used without significant loss of power.

In some applications, estimation of the treatment contrasts by simultaneous confidence intervals may be of greater interest. However, in many applications (such as in providing justification for the use of a new treatment over existing standards) hypothesis testing is used; here, stepwise multiple testing methods provide distinct power advantages over the single-step tests arising from applying methods intended primarily for estimation. The problem of deriving confidence intervals from stepwise tests is largely unresolved: see Hayter and Hsu (1994).

The results of this paper can be extended to two-sided tests in a straightforward manner. A Fortran program for computing the critical constants for two-sided tests is available from the authors. Unbalanced designs [involving unequal $Var\ \hat{\theta}_i$ and unequal $Corr\ (\hat{\theta}_i, \hat{\theta}_j)$] pose a more difficult problem, but the method of Dunnett and Tamhane (1991, 1995) developed for the step-down and step-up procedures can be used to implement SUDP(r) in this case. Although the resulting procedure cannot always be shown to control the familywise error rate, the excess over the nominal level α is usually quite small. Finally, we note that the adjusted p-values [see, e.g. Dunnett and Tamhane (1991, 1992b) and Westfall and Young (1993)] can also be defined and calculated for SUDP(r).

REFERENCES

Bechhofer, R.E., and Dunnett, C.W. (1988). Tables of percentage points of multivariate Student t distributions. *Sel. Tables Math. Statist.*, 11, 1–371.

Bechhofer, R.E., and Tamhane, A.C. (1981). Incomplete block designs for comparing treatments with a control. *Technometrics*, 23, 45–57.

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

Dunnett, C.W. (1989). Multivariate normal probability integrals with product correlation structure, Algorithm AS251. *Appl. Statist.*, 38, 564–579; Correction note, 42, 709. Program listing available from `http://lib.stat.cmu.edu/apstat/251` web address.

Dunnett, C.W. (1997). Comparisons with a control. *Encyclopedia of Statistical Sciences, Update Volume* 1, (S. Kotz, B.C. Read and D.L. Banks, *eds.*), Wiley, New York, 126–134.

Dunnett, C.W., and Tamhane, A.C. (1991). Step-down multiple tests for comparing treatments with a control in unbalanced one-way layouts. *Statist. Med.*, 10, 939–947.

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

Dunnett, C.W., and Tamhane, A.C. (1992b). Comparisons between a new drug and active and placebo controls in an efficacy clinical trial. *Statist. Med.*, 11, 1057–1063.

Dunnett, C.W., and Tamhane, A.C. (1993). Power comparisons of some step-up multiple test procedures. *Statist. Probab. Lett.*, 16, 55–58.

Dunnett, C.W., and Tamhane, A.C. (1995). Step-up multiple testing of parameters with unequally correlated estimates. *Biometrics*, 51, 217–227.

Hayter, A.J., and Hsu, J.C. (1994). On the relationship between stepwise decision procedures and confidence sets. *J. Amer. Statist. Assoc.*, 89, 128–136.

Hochberg, Y., and Tamhane, A.C. (1987), *Multiple Comparison Procedures*. Wiley, New York.

Laska, E.M., and Meisner, M.J. (1989). Testing whether an identified treatment is best. *Biometrics*, 45, 1139–1151.

Marcus, R., Gabriel, K.R., and Peritz, E. (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika*, 63, 655–660.

Miller, R.G., Jr. (1966). *Simultaneous Statistical Inference*. McGraw-Hill, New York.

Naik, U.D. (1975). Some selection rules for comparing p processes with a standard. *Comm. Statist. Ser. A*, 4, 519–535.

Westfall, P.H., and Young, S.S. (1993). *Resampling Based Multiple Testing*. Wiley, New York.

Received 21 February 1996
Revised 26 February 1997
Accepted 14 January 1998

Department of Statistics and
Department of Industrial Engineering and Management Sciences
Northwestern University
Evanston, Illinois
U.S.A. 60208
e-mail: ajit@iems.nwu.edu

Department of Mathematics
University of Southampton
Southampton, SO17 1BJ
UK

Department of Mathematics and Statistics and
Department of Clinical Epidemiology and Biostatistics
McMaster University
Hamilton, Ontario
Canada L8S 4K1
email: math@mcmail.mcmaster.ca