ELSEVIER

Journal of Statistical Planning and Inference 82 (1999) 55–68

journal of statistical planning and inference

www.elsevier.com/locate/jspi

Stepwise multiple test procedures with biometric applications

Ajit C. Tamhane[a], *, Charles W. Dunnett[b], [1]

[a] *Department of Statistics, Northwestern University, Evanston, IL 60208, USA*

[b] *Department of Mathematics and Statistics, McMaster University, Hamilton, Ont, Canada L8S 4K1*

Received 1 January 1997; accepted 1 October 1997

Abstract

We provide an overview of the recent developments in normal theory stepwise multiple test procedures for non-hierarchical families and describe several biometric applications where these procedures are useful. © 1999 Elsevier Science B.V. All rights reserved.

MSC: 62J15; 62P10

Keywords: Multiple comparisons; Stepwise test procedures; Familywise error rate; Biometric applications

1. Introduction

In this paper we provide an overview of normal theory multiple test procedures (MTPs) based on Student t statistics and describe some of their biometric applications. Normal theory MTPs have the advantage that they take into account the joint distribution of the test statistics. There are analogs of these procedures based on p-values associated with the individual test statistics which make no assumptions regarding the joint distribution or the form of the test statistics. These procedures may be used when the standard normal theory assumptions are not met and the test statistics are arbitrary. We will mention these procedures, but will not discuss them in detail; see Tamhane (1996) for a review of both normal theory and p-value-based procedures.

We describe normal theory MTPs for balanced designs in Section 3 and for unbalanced designs in Section 4. Biometric applications of the stepwise MTPs are described

* Corresponding author. Tel.: +1-708-491-3577; fax: +1-708-491-8005.

E-mail address: ajit@lagrange.iems.nwu.edu (A.C. Tamhane)

[1] Supported by a grant from the Natural Sciences and Engineering Research Council of Canada.

PII: S0378-3758(99)00031-2

in Section 5. Some open problems and directions for future research are discussed in Section 6.

It should be noted that this is a review paper and almost all of the procedures described herein either have been published elsewhere or are in the publication process. Because of this reason all mathematical proofs are omitted; the interested reader can refer to source papers for these and additional details.

2. Preliminaries

Consider $k \geqslant 2$ null hypotheses, $\mathrm{H}_1, \mathrm{H}_2, \ldots, \mathrm{H}_k$. We assume that the hypotheses form a non-hierarchical family, i.e., no H_i implies any other H_j. The familywise error rate (FWE) of an MTP is the probability that it rejects at least one true H_i. For a prespecified significance level α, an MTP is required to control its

$$\mathrm{FWE} \leqslant \alpha, \tag{2.1}$$

under all partial null hypotheses $\mathrm{H}_I = \bigcap_{i \in I} \mathrm{H}_i$ where I is any non-empty subset of the index set $\{1, 2, \ldots, k\}$. This is called strong control of the FWE (Hochberg and Tamhane, 1987). Strong control of the FWE is needed when we want to control the probability of making any type I error, no matter how many of the k hypotheses are true and how many are false. An MTP satisfying the strong control condition will be referred to as an α-level MTP.

A useful concept is the multiplicity adjusted p-value for H_i, denoted by $p_{\mathrm{a}i}$. It is the largest significance level at which H_i can be rejected for given data using a given MTP. Once the $p_{\mathrm{a}i}$ are computed, the MTP can be applied at a specified level α by rejecting any hypothesis H_i with $p_{\mathrm{a}i} < \alpha$.

MTPs can be divided into two broad categories: single-step procedures (SSPs) and stepwise procedures (SWPs). SWPs can be further subdivided into step-down procedures (SDPs) and step-up procedures (SUPs). In an SSP the decision about any hypothesis H_i does not depend on the decision about any other hypothesis H_j; therefore the hypotheses can be tested without reference to one another. In an SWP, on the other hand, the hypotheses are tested in a specific order, generally determined by the magnitudes of the test statistics or the associated p-values, p_i, and the decisions on them are made in a stepwise manner. The decisions on the earlier hypotheses in the order may affect those on the later hypotheses in the order. In an SDP the hypotheses are tested beginning with the most significant one and testing continues until a hypothesis is not rejected ('accepted'), at which point all the remaining hypotheses are accepted by implication without actually testing them. In an SUP, on the other hand, the hypotheses are tested beginning with the least significant one and testing continues until a hypothesis is rejected at which point all the remaining hypotheses are rejected by implication without actually testing them. In certain problems, e.g., dose finding (see Section 5.3), the order of the hypotheses for testing may be prespecified rather than determined by the magnitudes of the test statistics.

3. Procedures for balanced designs

3.1. Distributional setup

Consider the standard normal theory general linear model setting with k estimable parametric functions (typically contrasts among the treatment means), $\theta_1,\dots,\theta_k$. Let $\hat\theta_1,\dots,\hat\theta_k$ be their least-squares estimates. By a balanced design we mean that the $\hat\theta_i$ have equal variances and equal correlations. Specifically, we assume that the $\hat\theta_i$ have a joint k-variate normal distribution with

$$E(\hat\theta_i)=\theta_i,\ \mathrm{var}(\hat\theta_i)=\tau^2\sigma^2 \text{ and } \mathrm{corr}(\hat\theta_i,\hat\theta_j)=\rho \quad \text{for all } i\neq j, \tag{3.1}$$

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 v degrees of freedom (d.f.) so that the corresponding random variable (r.v.) S^2 is distributed as $\sigma^2\chi^2_v/v$ independently of the $\hat\theta_i$.

Three examples of this setup are: (i) 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; (ii) orthogonal contrasts among the cell means corresponding to main effects and interactions in a two-level factorial experiment with equireplicated cells; and (iii) a BTIB design (Bechhofer and Tamhane, 1981) for comparing treatments with a control using incomplete blocks.

We consider the following one-sided multiple hypotheses testing problem (the MTPs for the two-sided testing problem are analogous and hence are not discussed here):

$$\mathrm{H}_i\colon \theta_i=0 \text{ vs. } A_i\colon \theta_i>0 \quad (1\leqslant i\leqslant k).$$

The test statistics used to test the H_i are given by

$$t_i=\frac{\hat\theta_i}{\mathrm{SE}(\hat\theta_i)}=\frac{\hat\theta_i}{s\tau} \quad (1\leqslant i\leqslant k).$$

The r.v.'s T_i corresponding to the observed statistics t_i individually have Student t distributions and jointly have a k-variate t-distribution with common correlation ρ and d.f. v. The subset of the T_i corresponding to the true H_i has a central t-distribution, while the complementary subset has a noncentral t-distribution. Denote by $t^{(\alpha)}_{k,v,\rho}$, the upper α equicoordinate critical point of a central k-variate t-distribution with common correlation ρ and d.f. v. Comprehensive tables of these critical constants are given in Bechhofer and Dunnett (1988).

3.2. Single-step procedure (SSP)

We consider an α-level SSP that rejects any H_i for which

$$t_i>t^{(\alpha)}_{k,v,\rho} \quad (1\leqslant i\leqslant k). \tag{3.2}$$

The adjusted p-values for this SSP are computed under the overall null hypothesis $\mathrm{H}_0 = \bigcap_{i=1}^{k} \mathrm{H}_i$ using the formula

$$p_{\mathrm{a}i} = P\left\{ \max_{1 \leqslant j \leqslant k} T_j \geqslant t_i \right\} \quad (1 \leqslant i \leqslant k),$$

where $T_1, T_2, \ldots, T_k$ have a central k-variate distribution with common correlation ρ and d.f. v.

This SSP has associated with it the following $100(1-\alpha)\%$ simultaneous lower confidence bounds on the θ_i:

$$\theta_i \geqslant \hat{\theta}_i - t_{k,v,\rho}^{(\alpha)} s\tau \quad (1 \leqslant i \leqslant k).$$

Thus another equivalent way of applying this SSP is to reject any H_i for which the lower confidence bound on θ_i is positive. These lower confidence bounds are a special case (for equal ρ) of the bounds given by Dunnett (1955); we shall use the more general bounds in Section 4.2.

The corresponding p-value-based SSP is the conservative Bonferroni procedure which rejects any H_i with $p_i < \alpha/k$. This is equivalent to replacing $t_{k,v,\rho}^{(\alpha)}$ in (3.2) by the univariate t upper α/k critical point, $t_v^{(\alpha/k)}$, which is larger.

3.3. Step-down procedure (SDP)

Marcus et al.'s (1976) closure method can be used to construct an α-level MTP as follows: Reject any H_i iff every intersection hypothesis $\mathrm{H}_J = \bigcap_{j \in J} \mathrm{H}_j$ containing H_i is rejected at level α. To apply this method, one needs an α-level test of every intersection hypothesis H_J. If we use Roy's (1953) union-intersection test of H_J which rejects at level α if

$$\max_{j \in J} t_j > t_{|J|,v,\rho}^{(\alpha)},$$

where $|J|$ denotes the cardinality of set J, the resulting MTP for making decisions on the individual hypotheses H_i can be applied in a step-down manner. This SDP was suggested earlier by Miller (1966, pp. 85–86) but without a proof or even a claim of its strong FWE control property. The steps in this SDP are as follows:

Step 0: Order the test statistics t_i: $t_{(1)} \leqslant t_{(2)} \leqslant \cdots \leqslant t_{(k)}$. Let $\mathrm{H}_{(1)}, \mathrm{H}_{(2)}, \ldots, \mathrm{H}_{(k)}$ be the corresponding hypotheses.

Step 1: Reject $\mathrm{H}_{(k)}$ if $t_{(k)} > t_{k,v,\rho}^{(\alpha)}$ and go to Step 2. Otherwise accept all hypotheses and stop testing.

Step 2: Reject $\mathrm{H}_{(k-1)}$ if $t_{(k-1)} > t_{k-1,v,\rho}^{(\alpha)}$ and go to Step 3. Otherwise accept $\mathrm{H}_{(k-1)}, \ldots, \mathrm{H}_{(1)}$ and stop testing, etc.

In general, reject $\mathrm{H}_{(i)}$ iff $t_{(j)} > t_{j,v,\rho}^{(\alpha)}$ for $j = k, k-1, \ldots, i$.

The adjusted p-value for an ordered hypothesis $\mathrm{H}_{(i)}$ is given by

$$p_{a(i)} = \max(p'_{(k)}, p'_{(k-1)}, \ldots, p'_{(i)}),$$

where

$$p'_{(m)} = P\{\max(T_1, \ldots, T_m) \geqslant t_{(m)}\}$$

and $T_1, T_2, \ldots, T_m$ have a joint m-variate central equicorrelated t-distribution with common correlation ρ and d.f. ν.

Holm's (1979) p-value-based SDP can be viewed as a Bonferroni approximation to this SDP.

3.4. Step-up procedure (SUP)

Dunnett and Tamhane (1992a) proposed a step-up MTP in which the sequence of testing is reversed from that of the SDP. The steps in this SUP are as follows:

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

Step 1: Accept $\mathrm{H}_{(1)}$ if $t_{(1)} \leqslant c_1$ and go to Step 2. Otherwise reject all H_i and stop testing.

Step 2: Accept $\mathrm{H}_{(2)}$ if $t_{(2)} \leqslant c_2$ and go to Step 3. Otherwise reject $\mathrm{H}_{(2)}, \ldots, \mathrm{H}_{(k)}$ and stop testing, etc.

In general, accept $\mathrm{H}_{(i)}$ iff $t_{(j)} \leqslant c_j$ for $j = 1, 2, \ldots, i$.

The critical constants $c_1 \leqslant c_2 \leqslant \cdots \leqslant c_k$ are the solutions to the following equations: Let $(T_1, T_2, \ldots, T_k)$ have a central k-variate t-distribution with common correlation ρ and d.f. ν. Let $T_{1,m} \leqslant T_{2,m} \leqslant \cdots \leqslant T_{m,m}$ be the ordered values of $T_1, T_2, \ldots, T_m$. Then

$$P\{T_{1,m} \leqslant c_1, \ldots, T_{m,m} \leqslant c_m\} = 1 - \alpha \quad (1 \leqslant m \leqslant k). \tag{3.3}$$

These equations can be solved recursively starting with $m = 1$ in which case $c_1 = t_\nu^{(\alpha)}$, the upper α critical point of univariate Student's t. An algorithm for solving (3.3) is given in Dunnett and Tamhane (1992a) where tables of the constants c_i are provided for selected values of k, ν, ρ and α.

Numerical evaluation of powers reported in Dunnett and Tamhane (1993) shows that if only a few hypotheses are false then the SDP is slightly more powerful than the SUP, whereas if most hypotheses are false then the SUP is moderately more powerful than the SDP.

The adjusted p-values for the SUP are calculated as follows: Set $c_i = t_{(i)}$ and find $c_1, \ldots, c_{i-1}$ such that

$$P\{T_{1,j} \leqslant c_1, \ldots, T_{j,j} \leqslant c_j\} = 1 - p'_{(i)} \quad \text{for } j = 1, \ldots, i.$$

Then

$$p_{a(i)} = \min(p'_{(1)}, p'_{(2)}, \ldots, p'_{(i)}).$$

Hochberg's (1988) p-value-based SUP can be viewed as a Bonferroni approximation to this SUP.

3.5. Step-up–down procedure (SUDP)

The SDP begins by testing $t_{(k)} = t_{\max}$. This so-called MAX test answers the question 'Can at least one hypothesis be rejected?' If the answer to this question turns out to be affirmative, then the SDP proceeds in a step-down manner to provide a further resolution of this question by identifying the 'rejectable' hypotheses. The SUP begins by testing $t_{(1)} = t_{\min}$. This so-called MIN test (Laska and Meisner, 1989) answers the question 'Can all hypotheses be rejected?' If the answer to this question turns out to be negative, then the SUP proceeds in a step-up manner to provide a further resolution of this question by identifying the 'acceptable' hypotheses.

A generalization of the SDP and SUP can be obtained by posing the question 'Can at least q hypotheses be rejected?' A test to answer this question can be based on the statistic $t_{(r)}$ where $r = k + 1 - q$. The stepwise extension of this test proceeds in a step-down or step-up manner depending on whether the result of the test is significant or not. We call the resulting stepwise procedure a step-up–down procedure (SUDP(r)), where r is a prespecified integer between 1 and k. This procedure is studied in Tamhane et al. (1998). The SUP and SDP are special cases of SUDP(r) for $q = k, r = 1$ and $q = 1, r = k$, respectively.

The steps in SUDP(r) are as follows:

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

Step 1 (a): If $t_{(r)} \leqslant c_r$ then accept $\mathrm{H}_{(1)}, \mathrm{H}_{(2)}, \ldots, \mathrm{H}_{(r)}$ and go to General Step (a).

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

General Step (a): Let $\mathrm{H}_{(m)}$ denote the last accepted hypothesis (at Step 1 (a), $m=r$). If $m = k$ then stop testing; otherwise proceed as in the SUP and test $\mathrm{H}_{(m+1)}$. If $t_{(m+1)} > c_{m+1}$ then reject $\mathrm{H}_{(m+1)}, \mathrm{H}_{(m+2)}, \ldots, \mathrm{H}_{(k)}$ and stop testing. If $t_{(m+1)} \leqslant c_{m+1}$ then accept $\mathrm{H}_{(m+1)}$. Set $m = m + 1$ and return to the beginning of this step.

General Step (b): Let $\mathrm{H}_{(m)}$ denote the last rejected hypothesis (at Step 1 (b), $m=r$). If $m = 1$ then stop testing; otherwise proceed as in the SDP and test $\mathrm{H}_{(m-1)}$. If $t_{(m-1)} \leqslant c_{m-1}$ then accept $\mathrm{H}_{(m-1)}, \mathrm{H}_{(m-2)}, \ldots, \mathrm{H}_{(1)}$ and stop testing. If $t_{(m-1)} > c_{m-1}$ then reject $\mathrm{H}_{(m-1)}$. Set $m = m - 1$ and return to the beginning of this step.

It is shown in Tamhane et al. (1998) that the critical constants of SUDP(r) which satisfy (2.1) are given by the following: For $m = 1, 2, \ldots, r$, choose $c_m = t^{(\alpha)}_{m,v,\rho}$, and for $m > r$, determine the c_m recursively from the equation

$$P(T_{r,m} \leqslant t^{(\alpha)}_{r,v,\rho}; T_{r+1,m} \leqslant c_{r+1}, \ldots, T_{m,m} \leqslant c_m) = 1 - \alpha,$$

where $T_{1,m} \leqslant T_{2,m} \leqslant \cdots \leqslant T_{m,m}$ are the order statistics of $T_1, T_2, \ldots, T_m$, which have a central m-variate t distribution with common correlation ρ and d.f. v. Table 1 gives the critical constants of SUDP(r) for $k = 5, v = \infty, \rho = 0.5, \alpha = 0.05$ and for $r = 1, 2, \ldots, 5$. Note that for $r = 1$ the critical constants coincide with those of the SUP and for $r = 5$ they coincide with those of the SDP.

Another way to view this generalization is as follows. Let us suppose that the actual numbers of true and false hypotheses, p and $q = k - p$, are known. Among all SUDP(r) for $r = 1, \ldots, k$, which procedure is the most powerful? As noted earlier, if q is small (e.g., $q = 1$) then the SDP is more powerful than the SUP, while if q is large (e.g., $q = k$) then the SUP is more powerful than the SDP. It can be shown that SUDP(r)

Table 1
Critical Constants for SUDP(r) for $k=5, \nu=\infty, \rho=0.5, \alpha=0.05$

Procedure	c_1	c_2	c_3	c_4	c_5
SUDP(1)	1.645	1.933	2.071	2.165	2.237
SUDP(2)	1.645	1.916	2.068	2.164	2.237
SUDP(3)	1.645	1.916	2.062	2.164	2.236
SUDP(4)	1.645	1.916	2.062	2.160	2.236
SUDP(5)	1.645	1.916	2.062	2.160	2.234

with $r=p+1=k+1-q$ is the most powerful procedure. Power studies by Tamhane et al. (1998) show that if $p>0$ then the SUP suffers only a small loss in power in comparison to the most powerful SUDP(r). On the other hand, if $p=0$ then the SUP $=$ SUDP(1) is the most powerful procedure and SUDP(2),$\ldots$,SUDP(k) suffer a larger loss in power in comparison to the SUP. Since p and q are unknown, the SUP may be regarded as the preferred procedure since it minimizes the maximum loss in power in comparison to the most powerful SUDP(r) for any $r=p+1$.

4. Procedures for unbalanced designs

4.1. Distributional setup

By an unbalanced design we mean that the distributional setup is the same as in Section 3.1 except that the $\hat{\theta}_i$ have possibly unequal variances and unequal correlations. In other words, instead of (3.1) we have

$$E(\hat{\theta}_i)=\theta_i,\ \mathrm{var}(\hat{\theta}_i)=\tau_i^2\sigma^2 \text{ and } \mathrm{corr}(\hat{\theta}_i,\hat{\theta}_j)=\rho_{ij} \quad \text{for all } i\neq j, \tag{4.1}$$

here τ_i^2 and ρ_{ij} are known design-dependent constants.

Example. A common example of this setup is the one-way layout for comparing treatments with a control in which the sample sizes of the treatment groups are not all equal. Specifically, let n_0 be the sample size of the control group and let n_i be the sample size of the ith treatment group $(1\leqslant i\leqslant k)$. Assume that the observations in the ith group are independent and normally distributed with mean μ_i and a common variance σ^2 $(0\leqslant i\leqslant k)$, and observations in different groups are independent of each other. Let $\bar{y}_i$ be the sample mean for the ith group and let s^2 be the mean-square error estimate of σ^2 with $\nu=\sum_{i=0}^{k} n_i-(k+1)$ d.f. The contrasts of interest are $\theta_i=\mu_i-\mu_0$ and their least-squares estimates are $\bar{y}_i-\bar{y}_0$ $(1\leqslant i\leqslant k)$. Then

$$\tau_i^2=\frac{1}{n_0}+\frac{1}{n_i} \quad \text{and} \quad \rho_{ij}=\sqrt{\frac{n_i}{n_0+n_i}}\sqrt{\frac{n_j}{n_0+n_j}} \quad (1\leqslant i\neq j\leqslant k). \tag{4.2}$$

The t statistics for testing the hypotheses H_i are given by

$$t_i=\frac{\hat{\theta}_i}{\mathrm{SE}(\hat{\theta}_i)}=\frac{\hat{\theta}_i}{s\tau_i} \quad (1\leqslant i\leqslant k).$$

4.2. Single-step procedure (SSP)

The SSP remains unchanged from its description in Section 3.2 except that instead of the upper α critical point, $t_{k,v,\rho}^{(\alpha)}$, of an equicorrelated (with common correlation ρ) k-variate t-distribution, we use the critical point of an unequally correlated k-variate t-distribution with correlation matrix $R_k = \{\rho_{ij}\}$, denoted by $t_{k,v,R_k}^{(\alpha)}$. The $100(1-\alpha)\%$ simultaneous confidence lower bounds on the θ_i associated with this SSP are given by

$$\theta_i \geqslant \hat{\theta}_i - t_{k,v,R_k}^{(\alpha)} s\tau_i \quad (1 \leqslant i \leqslant k).$$

Although the correlations are unequal, it is not unduly burdensome to compute $t_{k,v,R_k}^{(\alpha)}$ using the program by Dunnett (1989) which exploits the product correlation structure indicated in (4.2). This algorithm calculates multivariate normal probabilities over rectangular regions. The Fortran code for this algorithm is available on Web site http://stat.lib.cmu.edu/apstat which also includes the code for doing the additional integration required to uncondition over S/σ. For arbitrary correlations, instead of Dunnett's algorithm, Schervish's (1984) algorithm can be used which is also available on the same Web site. However, it requires much more computing time unless k is small. A very good approximation to $t_{k,v,R_k}^{(\alpha)}$ can be obtained by replacing the ρ_{ij} by their arithmetic average.

4.3. Step-down procedure (SDP)

Dunnett and Tamhane (1991) proposed that the SDP be carried out as before except that for testing $t_{(m)}$ use the critical constant $t_{m,v,R_m}^{(\alpha)}$, which is the upper α point of an m-variate central t-distribution with v d.f. and correlation matrix R_m corresponding to the m smallest t-statistics ($m = k, k-1, \ldots, 1$). In the above example of comparisons with a control in an unbalanced one-way layout, if $n_{(1)}, n_{(2)}, \ldots, n_{(k)}$ denote the sample sizes associated with the treatment groups yielding the ordered statistics $t_{(1)} \leqslant t_{(2)} \leqslant \cdots \leqslant t_{(k)}$, respectively, then the (i,j)th entry of R_m is

$$\rho_{ij} = \sqrt{\frac{n_{(i)}}{n_0 + n_{(i)}}} \sqrt{\frac{n_{(j)}}{n_0 + n_{(j)}}}.$$

Note that the critical constants used by this modified SDP depend on the observed ordering of the t-statistics. Conditional on this observed ordering, the r.v.'s associated with $t_{(1)}, \ldots, t_{(m)}$ are not jointly t-distributed. As a result, the choice of $t_{m,v,R_m}^{(\alpha)}$ for critical constants may not control the FWE, although the simulations performed by Dunnett and Tamhane (1995) suggested that it does. Liu (1996) pointed out that since this ordering is random, the critical constants should be determined by taking the maximum of the c_m-values for each m over all possible orderings, $t_{i_1} \leqslant t_{i_2} \leqslant \cdots \leqslant t_{i_k}$, of the t-statistics, where c_m denotes the critical constant for testing $t_{(m)}$. In other words, if $\{i_1, i_2, \ldots, i_m\}$ denotes any subset of $\{1, 2, \ldots, k\}$ then c_m must satisfy

$$\min_{1 \leqslant i_1 < \cdots < i_m \leqslant k} P\{\max(T_{i_1}, \ldots, T_{i_m}) \leqslant c_m\} = 1 - \alpha \quad (1 \leqslant m \leqslant k),$$

where $T_{i_1},\dots,T_{i_m}$ have a central m-variate t-distribution with v d.f.; the correlation matrix of this distribution is the submatrix corresponding to the indices $i_1, i_2,\dots,i_m$ of the $k \times k$ correlation matrix of $T_1, T_2,\dots,T_k$. Clearly, $c_1 = t_v^{(\alpha)}$, but for $m > 1$, the determination of the c_m-values using this approach presents a formidable computational task. However, for the above example of comparisons with a control in an unbalanced one-way layout, Liu (1996) showed that, in order to guarantee the FWE requirement (2.1), the least favorable ordering (in terms of maximizing c_m for $m = 1, 2,\dots,k$) is obtained when the ordered t-statistics, $t_{i_1} \leqslant t_{i_2} \leqslant \cdots \leqslant t_{i_k}$, correspond directly with the ordered sample sizes, $n_{i_1} \leqslant n_{i_2} \leqslant \cdots \leqslant n_{i_k}$. Therefore, the critical constants required to control the FWE should be determined by assuming the n's in the order $n_1 \leqslant n_2 \leqslant \cdots \leqslant n_k$ (which is the least favorable case) instead of the order associated with the ordered t-statistics.

Although it is possible that the modification of the SDP proposed in Dunnett and Tamhane (1991) for unbalanced designs may fail to control the FWE in some cases, our simulations indicate that any excess in FWE over the nominal value is likely to be quite small. On the other hand, Liu's (1996) modification tends to be conservative in general, besides being computationally burdensome. Therefore, if the possibility of small excesses in the FWE can be tolerated then our modification may be a preferred practical alternative.

4.4. Step-up procedure (SUP)

Dunnett and Tamhane (1995) proposed a modification of the SUP for unbalanced designs analogous to their modification of the SDP, which involves computation of the critical constants based on the observed ordering of the t-statistics. This modified SUP suffers from the same drawback in that it may not control the FWE in all cases. Grechanovsky and Pinsker (1996) have constructed some counterexamples to demonstrate this fact.

Liu (1996) proposed the same approach as that for the SDP to determine the critical constants of the SUP for unbalanced designs, i.e., find the maximum of the critical constants to control the FWE over all possible orderings of the t-statistics. Unfortunately, this computational problem does not simplify as it does in the case of the SDP for the example of comparisons with a control in an unbalanced one-way layout. In particular, the ordering of the t_i, $t_{i_1} \leqslant t_{i_2} \leqslant \cdots \leqslant t_{i_k}$, corresponding to the ordered sample sizes, $n_{i_1} \leqslant n_{i_2} \leqslant \cdots \leqslant n_{i_k}$, is not necessarily least favorable for the SUP. Our recommendation therefore is to use Dunnett and Tamhane's (1995) modified SUP since it is only in exceptional cases that it fails to control the FWE, and that too by fairly small amounts.

4.5. Step-up–down procedure (SUDP)

The critical constants of SUDP(r) can be calculated by employing similar methods to those used for the SDP in Dunnett and Tamhane (1991) and for the SUP in Dunnett and Tamhane (1995). As noted before, these methods do not guarantee control of the FWE under all configurations, but any excesses over the nominal value α are likely to be quite small. Suppose that the hypotheses are labelled so that $t_1 \leqslant t_2 \leqslant \cdots \leqslant t_k$. Let $T_1, T_2, \dots, T_k$ be the corresponding r.v.'s; note that the T_i are not ordered. Let R_m be the correlation matrix corresponding to the r.v.'s $T_1, T_2, \dots, T_m$. Then for fixed r, the critical constants, c_m for $m = 1, 2, \dots, r$ are given by $c_m = t^{(\alpha)}_{m,\nu,R_m}$ where $t^{(\alpha)}_{m,\nu,R_m}$ is the upper α equicoordinate critical point of the central m-variate t distribution with ν d.f. and correlation matrix R_m. For $m > r$, the critical constants can be found recursively from the equation

$$P(T_{r,m} \leqslant t^{(\alpha)}_{r,\nu,R_m}; T_{r+1,m} \leqslant c_{r+1}, \dots, T_{m,m} \leqslant c_m) = 1 - \alpha,$$

where $T_{1,m} \leqslant T_{2,m} \leqslant \cdots \leqslant T_{m,m}$ are the order statistics of $T_1, T_2, \dots, T_m$, which have a central m-variate t distribution with ν d.f. and correlation matrix R_m.

5. Biometric applications

5.1. Comparisons of treatments with a control

As alluded to earlier, the most common application of the MTPs discussed in this paper is to the classical problem of comparisons of test treatments with a control to determine which treatments are more effective than the control, and to select one of them. Let μ_0 denote the mean response of the control and μ_i that of the ith treatment. The hypotheses to be tested are

$$\mathrm{H}_i\colon \mu_i - \mu_0 \leqslant 0 \text{ vs. } A_i\colon \mu_i - \mu_0 > 0 \quad (1 \leqslant i \leqslant k), \tag{5.1}$$

assuming that a larger response indicates higher efficacy. We will assume the one-way layout setting of the example in Section 4, in which case the τ_i and the ρ_{ij} are known quantities, being functions of the group sample sizes as given in (4.2). The SSP may be used in this case if simultaneous confidence bounds are desired on all $\mu_i - \mu_0$. Otherwise the SDP or SUP are more powerful alternatives, with the SUP being the preferred choice if there is no prior knowledge about how many hypotheses are false.

D'Agostino and Heeren (1991) discussed the efficacy evaluation of an over-the-counter test drug where the testing problem (5.1) arises in a different context. The first objective in the efficacy evaluation is to demonstrate the sensitivity of the study by checking whether the study can detect that $k \geqslant 2$ *known* active treatments (standard drugs) are more effective than the placebo control. D'Agostino and Heeren stipulated that all the active treatments must be shown to be effective (i.e., all the H_i must be rejected) in order for the study to be regarded as sensitive and suggested the use of the SSP; Dunnett and Tamhane (1992b) as well as some discussants of the D'Agostino–Heeren paper pointed out that the correct test to use for this stipulation is the MIN test of Laska and Meisner (1989). If the MIN test fails to reject all hypotheses, the active treatments which fail to show efficacy may be identified using the SUP. One could then investigate the reasons for their failure which could be losses in sample sizes or lack of compliance with the protocol. If the stipulated requirement for the study to be regarded as sensitive is modified as at least q out of k active treatments be shown effective then SUDP(r) with $r = k + 1 - q$ may be used.

After the sensitivity of the study is established, the next step is to demonstrate the efficacy of the test drug with respect to the placebo control. This is a single test. Finally, once the efficacy of the test drug is demonstrated for its regulatory approval, it may be of interest to the sponsor for marketing purposes to determine if their test drug is more effective than any of the standard drugs. Here the appropriate MTP to use is the SDP or the SUP depending on the sponsor's prior expectation of whether the test drug would be more effective than only a few standards or most standards, respectively. If simultaneous confidence intervals on the efficacy differences between the test and standard drugs are desired then the SSP must be used. Dunnett and Tamhane (1992b) argued that, although three different testing problems are addressed in this efficacy evaluation, a separate familywise α level may be used for each.

Another related problem is that of the evaluation of a combination drug by comparing it to all of its subcombinations. Here the combination drug is the control with μ_0 being its mean response and μ_i is the mean response of the ith subcombination. The hypotheses are reversed from (5.1): they are H$_i$: $\mu_i - \mu_0 \geqslant 0$ vs. A_i: $\mu_i - \mu_0 < 0$ $(1 \leqslant i \leqslant k)$. If the combination drug must be shown to be more effective than *all* of its subcombinations in order for it to be acceptable then the MIN test may be used. If the requirement is relaxed to allow for the combination to be more effective than at least q out of k subcombinations then SUDP(r) with $r = k + 1 - q$ may be used.

5.2. Dose finding

A typical dose-finding problem involves comparing $k \geqslant 2$ increasing dose levels of a chemical compound and a zero dose level with respect to a certain response. Label the dose levels $0, 1, \ldots, k$ and let $\mu_0, \mu_1, \ldots, \mu_k$ denote the corresponding mean responses. For simplicity, consider a balanced one-way layout with an equal number, n, of observations at each dose level (including the zero dose level). Denote by $\bar{y}_0, \bar{y}_1, \ldots, \bar{y}_k$ the sample means and by s^2 the mean square error estimate of the experimental error variance, σ^2, with $v = (k + 1)(n - 1)$ d.f.

Assume that the μ_i are ordered:

$$\mu_0 \leqslant \mu_1 \leqslant \cdots \leqslant \mu_k. \tag{5.2}$$

The goal is to find the lowest dose level for which the response μ_i exceeds μ_0 by a prespecified threshold amount, $\delta \geqslant 0$. Without loss of generality we take $\delta = 0$. When the response is the efficacy of the compound, this dose is referred to as the minimum effective dose (MED) defined as

$$\text{MED} = \min\{i \colon \mu_i > \mu_0\}.$$

Following Ruberg (1989), Tamhane et al. (1996) formulated the problem of identifying the MED as the following multiple hypotheses testing problem:

$$\mathrm{H}_i\text{: } \mu_0 = \mu_1 = \cdots = \mu_i \text{ vs. } A_i\text{: } \mu_0 < \mu_i \quad (1 \leqslant i \leqslant k). \tag{5.3}$$

Note that the hypotheses H_i form a closed family since H_i implies H_j if $i > j$. These hypotheses can be tested either in a step-up or step-down manner. The estimated MED (sometimes referred to as the minimum detectable dose or the MDD) is given by the first rejected hypothesis when using a step-up MTP and by the last rejected hypothesis when using a step-down MTP.

A variety of test statistics can be used to test the hypotheses. One class of statistics is based on testing the significance of contrasts

$$C_i = \sum_{j=0}^{k} c_{ij}\bar{y}_j \quad (1 \leqslant i \leqslant k),$$

using the t-statistics

$$t_i = \frac{C_i}{\mathrm{SE}(C_i)} = \frac{C_i}{s\sqrt{1/n\sum_{j=0}^{k} c_{ij}^2}} \quad (1 \leqslant i \leqslant k),$$

here $\sum_{j=0}^{k} c_{ij} = 0$ for each set of contrast coefficients, $\{c_{i0}, c_{i1}, \ldots, c_{ik}\}$. Sets of contrast coefficients can be chosen in different ways. Which choice gives a more powerful MTP depends on the unknown shape of the dose response function and the true value of the MED.

The simulations performed by Tamhane et al. (1996) showed that for the types of dose response functions considered, SDPs are generally at least as powerful as SUPs. Two SDPs were considered for each set of contrasts: SDP1 was the same as the SDP of Section 3.3 which used the statistic $\max_{1 \leqslant i \leqslant m} t_i$ for testing H_m with the critical value $t_{m,v,1/2}^{(\alpha)}$ for $m = k, k-1, \ldots,$ stopping the first time a hypothesis is not rejected. SDP2 used the statistic t_i for testing H_i comparing it with the univariate t upper α point, $t_v^{(\alpha)}$, stopping the first time a hypothesis is not rejected. SDP1 was generally found to be more powerful.

Dunnett and Tamhane (1998) studied additional SDPs for the above dose-finding problem. Using simulations they found that SDP2 based on Bartholomew's (1959a, b) test generally has the highest power. Unfortunately, its critical points are difficult to obtain for unbalanced designs. A simpler alternative that achieves almost the same powers is another SDP based on the so-called step contrasts. Specifically, beginning with $m = k$, this SDP tests H_m by comparing the average of $\bar{y}_{m-r+1}, \ldots, \bar{y}_m$ with the average of $\bar{y}_0, \bar{y}_1, \ldots, \bar{y}_{m-r}$ for $r = 1, \ldots, m$, and rejects H_m if the maximum of the m t-statistics for these contrasts exceeds the upper α point of the m-variate central t-distribution with the correlation matrix determined by the contrast coefficients. Testing continues in this manner as long as rejection occurs, reducing m by one at each step.

Bauer (1997) showed that only the pairwise contrasts: $C_i = \bar{y}_i - \bar{y}_0$, when used in SDP1 and SDP2 control the type I FWE regardless of whether the monotonicity condition (5.2) is satisfied or not. All other contrasts can lead to excessive type I and type II error probabilities if that condition is not satisfied. However, pairwise contrasts do not exploit any prior knowledge about the shape of the dose response function, and hence are not very powerful. This dilemma is not yet resolved.

6. Concluding remarks

As seen in this review, considerable progress has been made in recent years in the development of stepwise MTPs, both normal theory based and p-value based. Many open problems still remain, however. A few of these are briefly discussed below in no particular order.

1. It is necessary to extend the normal theory procedures to the case of unequal group variances. A Welch–Satterthwaite-type approximation can be used for the distributions of the test statistics, but the details of the methods need to be worked out and the accuracy of the approximations with regard to control of the FWE needs to be checked.
2. It is also necessary to extend the normal theory procedures to the case of unknown and unequal correlations between the test statistics. This extension will be useful for dealing with the problem of comparing a treatment group with a control group on multiple endpoints.
3. In many biometric studies the response is binary (success/failure) and the parameter of interest is the probability of success. It would be of a great practical value to develop appropriate MTPs for the binomial distribution model which may be used in this case. The work of Neuhäuser and Hothorn (1997) should prove very useful for this extension. The p-value-based MTPs, although applicable, may not be very powerful because they do not exploit the information about the joint distributions of the test statistics.
4. In some toxicology dose-response studies, the response is ordinal rather than numerical, e.g., the scale $0, +, ++, +++$ used for pathological findings. A suitable model needs to be developed for formulating the dose-finding problem in this setting and appropriate MTPs need to be developed.

It is hoped that the present review will give an impetus to the researchers to investigate these and other problems.

References

Bartholomew, D.J., 1959a. A test of homogeneity for ordered alternatives. Biometrika 46, 36–48.

Bartholomew, D.J., 1959b. A test of homogeneity for ordered alternatives. Biometrika 46, 328–335.

Bauer, P., 1997. A note on multiple test procedures for dose finding. Biometrics 53, 1125–1128.

Bechhofer, R.E., Dunnett, C.W., 1988. Tables of percentage points of multivariate Student t distributions. Selected Tables Math. Statist. 11, 1–371.

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

D'Agostino, R.B., Heeren, T.C., 1991. Multiple comparisons in over-the-counter drug clinical trials with both positive and placebo controls (with comments and rejoinder). Statist. Med. 10, 1–31.

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.

Dunnett, C.W., 1997. Comparisons with a control, in: Kotz, S., Read, C.B., Banks, D.L. (Eds.), Encyclopedia of Statistical Sciences, Update vol. 1. Wiley, New York, pp. 126–134.

Dunnett, C.W., Tamhane, A.C., 1991. Step-down multiple tests for comparing treatments with a control in unbalanced one-way layouts. Statist. Med. 11, 1057–1063.

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

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

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

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

Dunnett, C.W., Tamhane, A.C., 1998. Some new multiple test procedures for dose finding. J. Biopharm. Stat. 8, 353–366.

Grechanovsky, E., Pinsker, I., 1996. A general approach to stepup multiple test procedures for free-combinations families. Talk presented at the First International Conference on Multiple Comparisons, Tel-Aviv, Israel.

Hochberg, Y., 1988. A sharper Bonferroni procedure for multiple tests of significance. Biometrika 75, 800–802.

Hochberg, Y., Tamhane, A.C., 1987. Multiple Comparison Procedures. Wiley, New York.

Holm, S., 1979. A simple sequentially rejective multiple test procedure. Scand. J. Statist. 6, 65–70.

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

Liu, W., 1996. Step-down and step-up tests for comparing treatments with a control in unbalanced one-way layouts. Unpublished manuscript.

Marcus, R., Gabriel, K.R., 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.

Neuhäuser, M., Hothorn, L.A., 1997. Trend tests for dichotomous endpoints with application to carcinogenicity studies. Drug Inform. J. 31, 463–470.

Roy, S.N., 1953. On a heuristic method of test construction and its use in multivariate analysis. Ann. Math. Statist. 24, 220–238.

Ruberg, S.J., 1989. Contrasts for identifying the minimum effective dose. J. Amer. Statist. Assoc. 84, 816–822.

Schervish, M., 1984. Multivariate normal probabilities with error bound, Algorithm AS 195. Appl. Statist. 33, 81–94; Correction Note, Appl. Statist. 34, 103–104.

Tamhane, A.C., 1996. Multiple comparisons, in: Ghosh, S., Rao, C.R. (Eds.), Handbook of Statistics, vol. 13. Elsevier Science, Amsterdam, pp. 587–630, Chapter 18.

Tamhane, A.C., Hochberg, Y., Dunnett, C.W., 1996. Multiple test procedures for dose finding. Biometrics 52, 21–37.

Tamhane, A.C., Liu, W., Dunnett, C.W., 1998. A generalized step-up–down multiple test procedure. Can. J. Stat. 26, 353–363.