Taylor & Francis
Taylor & Francis Group

Multistage Tests of Multiple Hypotheses

JAY BARTROFF[1] AND TZE LEUNG LAI[2]

[1]Department of Mathematics, University of Southern California,
Los Angeles, California, USA
[2]Department of Statistics, Stanford University, Stanford, California, USA

Conventional multiple hypothesis tests use step-up, step-down, or closed testing methods to control the overall error rates. We will discuss marrying these methods with adaptive multistage sampling rules and stopping rules to perform efficient multiple hypothesis testing in sequential experimental designs. The result is a multistage step-down procedure that adaptively tests multiple hypotheses while preserving the family-wise error rate and extends Holm's (1979) step-down procedure to the sequential setting, yielding substantial savings in sample size with small loss in power.

1. Introduction

Multiple testing problems of the sequential or, more generally, multistage nature occur frequently in statistics. For example, in sequential fault detection and diagnosis (Lai, 2000; Nikiforov, 1995), after detecting that a change in the system has occurred at some time point, the task is to isolate this changepoint to one of k time intervals or diagnose it as one of k change types. Another area rich with examples is sequential clinical trials with multiple endpoints (e.g., Tang and Geller, 1999; Tang et al., 1993) in which patients are accrued, treated, and evaluated sequentially with regard to k different features, none of which may have priority. Some recent new areas of applications are quantitative finance, in empirical tests of the profitability of trading strategies (Romano and Wolf, 2005b), and genomics (Ge et al., 2003).

Based on data coming from a parametric family F_θ, $\theta \in \Theta$, of distributions, we will be concerned with testing a set of hypotheses $H_1, \ldots, H_k \subseteq \Theta$. A hypothesis H_i is *true* if the true θ lies in H_i. If $I \subseteq \{1, \ldots, k\}$ is the set of indices of the true

Address correspondence to Jay Bartroff, Department of Mathematics, University of Southern California, 3620 South Vermont Ave., KAP 108, Los Angeles, CA 90089-2532, USA; E-mail: bartroff@usc.edu

hypotheses, then the *family-wise error rate* (*FWE*) is defined as the probability

$$P(\text{some } H_i, i \in I, \text{ is rejected}). \tag{1.1}$$

Other and somewhat less stringent notions of error rate have been proposed (cf. Hochberg and Tamhane, 1987) but we focus here on FWE, the bounding of which is sometimes called *strong error control*, because of its prominence in clinical research (Lehmacher et al., 1991) and other applications.

A set of hypotheses $H_1, \ldots, H_k$ is called *closed* if the set $\{H_1, \ldots, H_k\}$ is closed under intersection. Marcus et al. (1976) introduced a method of testing a closed set of hypotheses $H_1, \ldots, H_k$ that controls the FWE by requiring that there be an α-level test of every intersection hypothesis $\bigcap_{i \in J} H_i$, $J \subseteq \{1, \ldots, k\}$. Let $\boldsymbol{H}$ be the set of all such intersections and note that closedness of $\{H_1, \ldots, H_k\}$ is equivalent to it being equal to $\boldsymbol{H}$. Beginning with the *global hypothesis* $\bigcap_{i=1}^k H_i$, Marcus et al.'s (1976) procedure tests the non-empty elements of $\boldsymbol{H}$ in order of decreasing *dimension* (defined as the maximum number of H_i being intersected) by their corresponding α-level tests and proceeds by using the rule that $H \in \boldsymbol{H}$ is tested if and only if all elements of $\boldsymbol{H}$ contained in H are tested and rejected. Multistage extension of closed tests is relatively straightforward because of the nature of the hypotheses and the test statistic. It is essentially a repeated closed testing procedure, performing closed testing of the hypotheses not yet rejected at every stage; see Tang et al. (1993) and Tang and Geller (1999).

When the set of hypotheses is not closed, Holm's (1979) step-down procedure is commonly used for fixed sample size problems. Beginning with a brief review of Holm's (1979) procedure, Sec. 2 then proceeds to provide a multistage extension, which aims to capture the generality of Holm's procedure for controlling the FWE and to be able to take advantage of the closed testing structure when it exists. A simulation study of the proposed procedure's power and expected sample size is given in Sec. 3, and further applications and relation to the existing litertaure are discussed in Sec. 4.

2. A Multistage Step-Down Procedure

Holm (1979) proposed the following general step-down method of testing $H_1, \ldots, H_k$ that does not assume closedness. Although Holm's procedure has been criticized for lack of power in some settings, it does preserve the FWE without making any assumptions about the structure of the hypotheses or correlations between the individual test statistics, only that for each hypothesis H_i there is a computable p-value $\hat{p}_i$ such that

$$P(\hat{p}_i \leq \alpha \,|\, H_i) \leq \alpha \tag{2.1}$$

for all $0 < \alpha < 1$. The α-level Holm's procedure proceeds as follows. Compute and order the p-values $\hat{p}_{i(1)} \leq \cdots \leq \hat{p}_{i(k)}$. For $j = 1, \ldots, k$, if

$$\hat{p}_{i(j)} \geq \alpha/(k - j + 1), \tag{2.2}$$

then accept $H_{i(j)}, \ldots, H_{i(k)}$; otherwise, reject $H_{i(j)}$ and move on to stage $j + 1$ (if $j < k$). A simple proof that the FWE of Holm's procedure is bounded by α is given in Lehmann and Romano (2005).

As noted above, when closedness exists it is unnecessary to use the step-down correction (2.2) or any Bonferroni-type correction, for that matter. However, it is illuminating to now consider how the step-down procedure is related to closed testing procedures. When closedness exists, Marcus et al.'s (1976) procedure can be viewed as a special case of Holm's step-down procedure in the following sense. Assume that the p-values "respect" the closedness in the sense that

$$H_i \subseteq H_j \Rightarrow \hat{p}_i \leq \hat{p}_j. \tag{2.3}$$

In this case Holm's procedure will test the elements of $\boldsymbol{H} = \{H_1, \ldots, H_k\}$ in order of decreasing dimension (provided we agree to use dimension to break any "ties"). Assuming that closedness and (2.3) hold, we make two slight modifications of Holm's procedure to utilize these properties. First, upon rejection of $H_{i(j)}$, we accept any of the remaining hypotheses whose complements are implied by $H_{i(j)}$; i.e., accept any $H_{i(j+1)}, \ldots, H_{i(k)}$ containing $H_{i(j)}^c$, the complement of $H_{i(j)}$. This does not change the FWE $\leq \alpha$ bound since (2.3) guarantees that the intersection of all true hypotheses, denoted by G, is the first true hypothesis tested. Otherwise, if G has already been accepted, then all true hypotheses are subsequently accepted according to this rule. Next, note that the Bonferroni-type correction in (2.2) is now unnecessary; i.e., the right-hand-side of (2.2) may be replaced by α while maintaining FWE $\leq \alpha$, since letting j_G denote the rank of the p-value associated with G,

$$\text{FWE} = P(G \text{ rejected})$$
$$\leq P(\hat{p}_{i(j_G)} < \alpha)$$
$$\leq \alpha.$$

We now introduce a multistage generalization of Holm's (1979) step-down procedure. As with the original version of Holm's procedure, we make no assumptions about the structure of the set of hypotheses $H_1, \ldots, H_k$ to be tested, and the only assumptions are about the family of available tests of the individual hypotheses through their significance levels, following the approach of Romano and Wolf (2005a). That this procedure satisfies FWE $\leq \alpha$ is proved in Theorem 2.1.

For each hypothesis H_i, assume there is a sequential test statistic $T_{i,n}$, a (non random) critical value function $C_n(\rho)$ that is non increasing in $\rho \in (0, 1)$, and a set N of possible sample sizes such that

$$\sup_{\theta \in H_i} P_\theta \left(\sup_{n \in N} [T_{i,n} - C_n(\rho)] \geq 0 \right) \leq \rho \tag{2.4}$$

for all $0 < \rho < 1$. This requirement is the multistage analog of (2.1). The existence of such a family of sequential test statistics may at first seem restrictive, but in many settings there are natural choices for the $T_{i,n}$; see the examples in Secs. 3 and 4. The use of the set N of possible sample sizes allows for the possibilities of fully sequential (e.g., $N = \{1, 2, 3, \ldots\}$) or group sequential (e.g., $N = \{m, 2m, 3m, 4m, 5m\}$ for some m) in both the truncated and non truncated settings.

The α-level multistage step-down procedure with no more than k stages is defined as follows. Let $I_1 = \{1, \ldots, k\}$, $n_0 = 0$, and let $|\cdot|$ denote set cardinality. For $j = 1, \ldots, k$:

1. Sample up to

$$n_j = \inf\left\{n \in N : n > n_{j-1} \text{ and } \max_{i \in I_j} T_{i,n} \geq C_n(\alpha/|I_j|)\right\}. \qquad (2.5)$$

2. Order the test statistics

$$T_{i(j,1),n_j} \geq T_{i(j,2),n_j} \geq \cdots \geq T_{i(j,|I_j|),n_j}.$$

3. Reject $H_{i(j,1)}, \ldots, H_{i(j,m_j)}$, where

$$m_j = \max\left\{m \geq 1 : \min_{1 \leq \ell \leq m}\left[T_{i(j,\ell),n_j} - C_{n_j}\left(\frac{\alpha}{|I_j| - \ell + 1}\right)\right] \geq 0\right\}. \qquad (2.6)$$

4. Stop if $j = k$, if $n_j = \sup N$, or if all remaining hypotheses contain the complement of some rejected hypothesis. Otherwise, let I_{j+1} be the indices of the remaining hypotheses and continue on to stage $j + 1$.

Theorem 2.1. *The multistage step-down procedure satisfies FWE $\leq \alpha$.*

Proof. Let I be the indices of the true hypotheses. If an error occurs, then for some $I_j \supseteq I$ and index ℓ such that $|I_j| - \ell + 1 \geq |I|$,

$$\max_{i \in I} T_{i,n_j} \geq C_{n_j}\left(\frac{\alpha}{|I_j| - \ell + 1}\right) \geq C_{n_j}(\alpha/|I|),$$

which implies that

$$\max_{i \in I} \sup_{n \in N}[T_{i,n} - C_n(\alpha/|I|)] \geq 0.$$

Then, using the Bonferroni inequality and (2.4),

$$\begin{aligned}
\text{FWE} &\leq P\left(\max_{i \in I} \sup_{n \in N}[T_{i,n} - C_n(\alpha/|I|)] \geq 0\right) \\
&\leq \sum_{i \in I} P\left(\sup_{n \in N}[T_{i,n} - C_n(\alpha/|I|)] \geq 0\right) \\
&\leq \sum_{i \in I} \alpha/|I| = \alpha.
\end{aligned}$$

$\square$

We point out additionally that the procedure may be stopped at Step 4 at any point as long as the remaining hypotheses are accepted, since this action can only serve to decrease the FWE. This feature may be of use in clinical trial applications; see the last paragraph of Sec. 4.2.

Theorem 2.1 holds regardless of the structure of the H_i or the joint distribution of the test statistics $T_{i,n}$. When the H_i are closed, the above multistage step-down procedure can be modified slightly to take advantage of this additional structure,

analogous to the discussion in the second paragraph of this section for the fixed-sample case. To this end, assume that the set $H_1, \ldots, H_k$ is closed and that the test statistics respect the closedness in the sense that

$$H_i \subseteq H_j \Rightarrow T_{i,n} \geq T_{j,n} \quad \text{for all } n \in N. \tag{2.7}$$

As in the fixed-sample case, this guarantees that the hypotheses are analyzed by decreasing dimension. First, we modify the multistage step-down procedure by changing Step 3 to:

3.′ For $\ell = 1, \ldots, m_j$: Reject $H_{i(j,\ell),n_j}$ and accept any remaining hypotheses containing $H_{i(j,\ell),n_j}^c$.

Next, we replace the fractions of α in (2.5)–(2.6) by α. These modifications do not cause violation of FWE $\leq \alpha$ by the same proof given in the second paragraph of this section, proving the following.

Theorem 2.2. *If the set of hypotheses $H_1, \ldots, H_k$ is closed and the test statistics $T_{i,n}$ satisfy* (2.7)*, then the multistage step-down procedure with Step* 3 *replaced by* 3′ *and α as the argument of C_n in* (2.5)–(2.6) *satisfies FWE $\leq \alpha$.*

3. Power and Expected Sample Size

The Holm step-down procedure's attractive quality is its generality; i.e., no assumptions about the structure of the hypotheses $H_1, \ldots, H_k$ or the individual test statistics, other than (2.1), are necessary. This generality is provided by the Bonferroni-type adjustment (2.2), which also can cause Holm's procedure to be conservative, in terms of FWE and power, relative to procedures that take into account correlations between the individual test statistics. This conservativeness is shared by the multistage step-down procedure because of its use of an analogous step-down rule (2.6). However, as pointed out above, the utility of either the multistage or fixed-sample step-down procedure lies in cases where such correlations are difficult to model.

Consider a multistage step-down procedure with maximum sample size $n = \max N$. Relative to the fixed-sample Holm step-down procedure of size n, the multistage procedure will have a reduction in expected sample size, provided the set N is chosen reasonably. But, by the Neyman-Pearson lemma, the power of the multistage procedure for rejecting a given false hypothesis cannot exceed the power of the fixed-sample Holm procedure. However, as shown by the following simulation study, this loss in power is usually slight while the savings in expected sample size tends to be substantial.

Consider a three-endpoint clinical trial where two of the endpoints concern continuous data and the third concerns probability of a certain binary outcome. For example, let the data be $X_i = (X_{i1}, X_{i2}, X_{i3})$, $i = 1, 2, \ldots$, where for $j = 1, 2$, the X_{ji} are i.i.d. normal random variables with unknown mean μ_j and variance 1, and the X_{3i} are independent Bernoulli random variables where $p = P(X_{3i} = 1)$ is unknown. Suppose the X_i represent clinical treatment outcomes for three endpoints of interest,

and it is desired to test efficacy of the treatment in the form of the three one-sided null hypotheses

$$H_1 : \mu_1 \le 0, \quad H_2 : \mu_2 \le 0, \quad H_3 : p \ge 1/2.$$

In cases such as this, the correlation between the components of X_i is likely to be unknown or difficult to model. In the following simulation study we compare the performance of the multistage step-down procedure with two other procedures in the three cases of independent, positively correlated and negatively correlated components of X_i. Table 1 contains the results for the independent case; the other two are discussed below. Whatever the correlation between the components, the three procedures evaluated are equally applicable since they do not depend on the correlation structure of the individual hypotheses. For Holm's step-down procedure, we use standard $\alpha = .05$-level likelihood ratio tests of H_1, H_2, H_3 of size $n = 35$ to have power around 90% at $(\mu_1, \mu_2, p) = (.5, .5, .75)$. For the multistage step-down procedure, we use the same test statistics in one-sided group sequential tests with $N = \{26, 29, 35\}$ and use a normal approximation for $\sum_i X_{3i}$ to compute $C_n(\rho)$ to satisfy (2.4). Table 1 contains the expected sample size, probability of rejecting each H_i, and FWE (when $\bigcup_{i=1}^{3} H_i$ is true) for the Holm procedure (denoted by H) and the multistage step-down procedure (denoted by MultH). To see the effects of the step-down rule (2.5)–(2.6), we also include the multistage test (denoted by Mult) identical to MultH but with α divided by $k = 3$ in place of the larger fractions of α in (2.5)–(2.6); see the first paragraph of Sec. 4.1. Each entry in Table 1 is computed

Table 1
A three-endpoint trial

Procedure	(μ_1, μ_2, p)	EM	$P(\text{rej. } H_1)$ (%)	$P(\text{rej. } H_2)$ (%)	$P(\text{rej. } H_3)$ (%)	FWE (%)
H		105	1.7	1.7	0.8	4.0
Mult	(0, 0, .5)	104.7	1.6	1.6	1.8	4.9
MultH		104.6	1.5	1.5	2.0	4.8
H		105	2.3	2.3	76.0	4.4
Mult	(0, 0, .75)	98.4	1.7	1.7	79.9	3.2
MultH		98.3	2.1	2.1	80.7	4.2
H		105	2.5	95.7	2.1	4.4
Mult	(0, .65, .5)	96.9	1.6	94.5	1.9	3.4
MultH		96.9	2.2	94.6	2.9	4.9
H		105	4.3	82.9	83.9	4.3
Mult	(0, .5, .75)	92.8	1.5	76.3	80.3	1.5
MultH		92.3	3.2	79.9	85.2	3.2
H		105	83.4	83.4	3.8	3.8
Mult	(.5, .5, .5)	93.5	76.3	76.3	1.8	1.8
MultH		93.1	79.6	79.6	2.7	2.7
H		105	70.9	70.9	86.8	NA
Mult	(.4, .4, .75)	89.9	55.3	55.3	80.2	NA
MultH		89.3	64.7	64.7	85.4	NA
H		105	88.6	88.6	90.0	NA
Mult	(.5, .5, .75)	87.2	76.4	76.4	80.0	NA
MultH		86.1	84.4	84.4	87.0	NA

from 50,000 simulation runs. The last six FWE entries are marked NA (not applicable) because none of the null hypotheses are true for those parameter values. The multistage procedures Mult and MultH show substantial savings in expected size sample over the Holm procedure's fixed sample size of 105 at significant deviations from the "worst-case" null $(\mu_1, \mu_2, p) = (0, 0, .5)$. The expected sample size of Mult and MultH are nearly identical, the former being somewhat larger due to its larger critical values in (2.5)–(2.6). For the same reason MultH has slightly higher power than Mult; in particular, see the last six rows of Table 1. Although the power of MultH is lower than H due to its multiple looks, this difference is slight, usually within a few percentage points. This relative relationship does not change when the components of X_i are correlated, it simply tends to decrease slightly in magnitude when positively correlated and increase when negatively correlated. For example, when $(\mu_1, \mu_2, p) = (0, .5, .75)$ and the two normal components have a correlation coefficient of .75, $P(\text{reject } H_1)$ increases to 4.5, 2.0, and 3.6% for H, Mult, and MultH, respectively, while the power $P(\text{reject } H_2)$ decreases to 80.7, 73.4, and 77.0%, respectively. Here Mult and MultH have expected sample size of 94.7 and 94.2.

4. Applications and Discussion

4.1. *Sequential k-Hypothesis Testing*

A straightforward application of (2.4) to sequential testing of k null hypotheses is to use Bonferroni's inequality so that the k hypotheses can be treated separately by setting $\rho = \alpha/k$ in (2.4). This approach to sequential multiple testing has been taken by many authors. Paulson (1964) noticed that further sample size savings might be possible by eliminating (rejecting) some hypotheses during the course of the experiment, similar to the test Mult in the preceding section. We have refined the multistage rejection of hypotheses by using Holm's step-down procedure to sharpen the Bonferroni bounds. In fact, when the set of hypotheses is closed, a slight modification of our multistage test can dispense with Bonferroni bounds, as shown in Theorem 2.2.

Sequential multiple hypothesis testing dated back to Sobel and Wald (1949) in deciding which of three simple hypotheses $H_1 : \theta = -d$, $H_2 : \theta = 0$, or $H_3 : \theta = d$ is true about a normal mean θ, for a fixed value $d > 0$. This is basically a classification problem. The Sobel–Wald test combines two sequential probability ratio tests (SPRTs) for different pairs of the three hypotheses, the comparison of H_1 versus H_3 being superfluous. Armitage (1950) generalized the Sobel–Wald problem to k hypotheses, corresponding to an error matrix $\alpha_{ij} = P_i(\hat{i} = j)$ for $i \neq j$, where P_i denotes the probability measure under H_i and $\hat{i}$ is the hypothesis chosen by the test. Armitage's test combines the corresponding $\binom{k}{2}$ SPRTs for the k hypotheses, which leads to stopping boundaries with slope $\pm 1/2$ in the $(n, S_n/d)$-plane, where $S_n = \sum_{i=1}^{n} X_i$ is the sum of the first n i.i.d. normal observations. Simons (1967) considered a generalization of a special case of Armitage's test for $k = 3$ in which the stopping boundaries' slopes can be chosen arbitrarily. Whereas the Sobel–Wald, Armitage, and Simons tests stop and decide on a hypothesis H_j when all the component SPRTs simultaneously prefer H_j to all other alternatives, Lorden (1972, 1976) introduced various multistage tests that decide on H_j only when all other hypotheses can be rejected, based on generalized likelihood ratios. Lorden's work was extended to

composite hypotheses by Pavlov (1988). Eisenberg (1991) gave a detailed summary of these problems.

Paulson (1963) introduced another generalization of the Sobel–Wald test, considering $k \geq 2$ intervals $(-\infty, \theta_1), (\theta_1, \theta_2), \ldots, (\theta_{k-1}, \infty)$ and testing sequentially to which interval θ belongs. In its symmetric form, Paulson's test stops the first time the interval (u_n, v_n) is contained in one of the intervals $(-\infty, \theta_1 + \delta), (\theta_{k-1} - \delta, \infty)$, or $(\theta_i - \delta, \theta_{i+1} + \delta), i = 1, \ldots, k-2$, where $\delta > 0$ is a chosen parameter and

$$u_n = \max_{1 \leq m \leq n} \tilde{u}_m = \max_{1 \leq m \leq n} (S_m/m - \delta/2 - A/m) \tag{4.1}$$

$$v_n = \min_{1 \leq m \leq n} \tilde{v}_m = \min_{1 \leq m \leq n} (S_m/m + \delta/2 + A/m) \tag{4.2}$$

in which $A > 0$ is a critical value that can be chosen to give desired coverage probability. Paulson's procedure can be viewed as a special case of the multistage step-down procedure, as follows. Defining $H_0^{(i)} : \theta = \theta_i - \delta$ and $H_1^{(i)} : \theta = \theta_i$ ($1 \leq i \leq k-1$), the one-sided SPRT of $H_0^{(i)}$ versus $H_1^{(i)}$ stops sampling and rejects $H_0^{(i)}$ if

$$S_n - n[(\theta_i - \delta) + \theta_i]/2 \geq A, \tag{4.3}$$

for some critical value $A > 0$. Dividing both sides of (4.3) by n and rearranging terms gives

$$A/n \leq S_n/n - (\theta_i - \delta/2) = S_n/n - \delta/2 - (\theta_i - \delta),$$

which by (4.1) is equivalent to $\tilde{u}_n \geq \theta_i - \delta$. Similarly, the one-sided SPRT of $\widetilde{H}_0^{(i)} : \theta = \theta + \delta$ versus $H_1^{(i)}$ stops sampling and rejects $\widetilde{H}_0^{(i)}$ if $\tilde{v}_n \leq \theta_i + \delta$. Hence, running the multistage step-down procedure until a coherent classification is made is precisely Paulson's procedure. In the small-probability event that no coherent classification is made—say, if at some stage a hypothesis is rejected containing the complement of a previously rejected hypothesis—classification based on S_n/n or randomization can be used, as Paulson (1963) suggested.

Although we have used $H_0^{(i)}$ and $\widetilde{H}_0^{(i)}$ to denote null hypotheses for combining SPRTs above, there is no natural notion of a "null" hypothesis in the actual classification problem. We could have populated our list of hypotheses to test in any way that suited our needs. Thus, it is also perhaps interesting to consider what closed testing has to say about the Sobel–Wald–Paulson problem. In particular, for the case $k = 3$ in the Paulson problem, let $H_1 : \theta < \theta_1$, $H_2 : \theta > \theta_1$, $H_3 : \theta < \theta_2$, and $H_4 : \theta > \theta_2$. The set $\{H_1, H_2, H_3, H_4\}$ is not closed, but adding $H_5 = H_2 \cap H_3 = (\theta_1, \theta_2)$ to the list of hypotheses completes its closure $\boldsymbol{H} = \{H_1, H_2, H_3, H_4, H_5\}$. Since all intersections of dimension 3 or higher are empty, the closed testing principle suggests beginning by testing the hypotheses of dimension 2, namely, $H_1 = H_1 \cap H_3 = (-\infty, \theta_1)$, $H_4 = H_2 \cap H_4 = (\theta_2, \infty)$, and $H_5 = (\theta_1, \theta_2)$, which is of course the original Paulson problem with $k = 3$. Hence, it seems that the closed testing principle does not give any new insight into the Paulson problem.

A closely related problem to sequential k-hypothesis testing is selecting the one of k normal populations with the largest mean. Bechhofer (1954) considered this problem when the variance of the observations is known and proposed a fixed sample procedure that compares the sample means of the individual populations. For unknown variance, Bechhofer et al. (1954) proposed a two-stage procedure,

and Robbins et al. (1968) proposed a sequential procedure with improved efficiency, also based on sample means. The special case of when the mean is an integer was considered by Robbins (1970) and later generalized by McCabe (1973). Robbins (1970) introduced the notion of "distinguishability" of a family of populations, and Khan (1973) studied the asymptotic efficiency of stopping rules that distinguish within such families. Mukhopadhyay (1983) proposed likelihood-based methods for the largest mean problem and used Khan's results to show asymptotic efficiency. Likelihood-based methods were shown to be useful in a number of related selection problems as well; see Mukhopadhyay and Solanky (1994). Further references are given in Sec. 2 of Chan and Lai (2005).

4.2. *Multiple Endpoint Clinical Trials*

The multistage step-down procedure provides a general method of testing multiple endpoints in clinical trials. The adaptive rejection times (2.5) and rejection rule (2.6) have the effect of adaptively "dropping" (i.e., rejecting) hypotheses when enough information has accumulated to do so, to focus on the statistically most interesting endpoints. As discussed above, when closedness exists, closed testing methods should be used since they are are in general more powerful than step-down methods (simulations verifying this were conducted by Lehmacher et al., 1991 and Tang and Lin, 1997) because they forgo the need for Bonferroni-type corrections, such as in (2.2), (2.5), and (2.6). The multistage step-down method could be useful in cases where closedness does not exist. For example, in clinical trials for AIDS treatments, it is common (e.g., Fischl et al., 1987) to have multiple endpoints of both the continuous and categorical types, like CD4 (T-cell) level, which is commonly modeled as a normal random variable, and the binary indicator of opportunistic infectious disease like a cold, modeled as a Bernoulli random variable. Moreover, if one or some subset of the endpoints is of primary interest, the multistage step-down procedure can be used as a pilot or screening phase with the option of immediately stopping and proceeding to secondary testing when one of the primary hypotheses is rejected. That this does not increase the FWE is pointed out following the proof of Theorem 2.1. The multistage step-down procedure provides a general framework that can be applied to multiple testing in these clinical trials.

Acknowledgments

References

Armitage, P. (1950). Sequential analysis with more than two alternative hypotheses, and its relation to discriminant function analysis. *J. Roy. Stat. Soc. B* 12:137–144.

Bechhofer, R. E. (1954). A single-sample multiple decision procedure for ranking means of normal populations with known variance. *Ann. Math. Stat.* 25:16–39.

Bechhofer, R. E., Dunnett, C. W., Sobel, M. (1954). A two-sample multiple decision procedure for ranking means of normal populations with a common unknown variance. *Biometrika* 41:170–176.

Chan, H. P., Lai, T. L. (2005). Importance sampling for generalized likelihood ratio procedures in sequential analysis. *Sequential Analysis* 24(3):259–278.

Eisenberg, B. (1991). Multihypothesis problems. In: Ghosh, B. K., Sen, P. K., eds. *Handbook of Sequential Analysis*. New York: Marcel Dekker, pp. 229–244

Fischl, M. A., et al. (1987). The efficiency of azidothymidine (AZT) in the treatment of patients with AIDS and AIDS-related complex. *New Eng. J. Med.* 317:185–191.

Ge, Y., Dudoit, S., Speed, T. P. (2003). Resampling-based multiple testing for microarray data analysis. *Test* 12:1–44.

Hochberg, Y., Tamhane, A. C. (1987). *Multiple Comparison Procedures*. New York: John Wiley & Sons.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scand. J. Stat.* 6:65–70.

Khan, R. A. (1973). On sequential distinguishability. *Ann. Stat.* 1:838–850.

Lai, T. L. (2000). Sequential multiple hypothesis testing and efficient fault detection-isolation in stochastic systems. *IEEE Trans. Inform. Theor.* 46:595–608.

Lehmacher, W., Wassmer, G., Reitmeir, P. (1991). Procedures for two-sample comparisons with multiple endpoints controlling the experimentwise error rate. *Biometrics* 47:511–521.

Lehmann, E. L., Romano, J. P. (2005). *Testing Statistical Hypotheses*. 3rd ed. New York: Springer.

Lorden, G. (1972). Likelihood ratio tests for sequential k-decision problems. *Ann. Math. Stat.* 43:1412–1427.

Lorden, G. (1976). 2-SPRT's and the modified Kiefer–Weiss problem of minimizing an expected sample size. *Ann. Stat.* 4:281–291.

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

McCabe, G. (1973). Estimation of the number of terms in a sum. *J. Am. Stat. Assoc.* 68:452–456.

Mukhopadhyay, N. (1983). Selecting the largest normal mean through likelihoods. *Stat. Decis.* 1:171–182.

Mukhopadhyay, N., Solanky, T. K. S. (1994). *Multistage Selection and Ranking Procedures: Second-Order Asymptotics*. New York: Marcel Dekker.

Nikiforov, I. V. (1995). A generalized change detection problem. *IEEE Trans. Inform. Theor.* 41:171–187.

Paulson, E. (1963). A sequential decision procedure for choosing one of k hypotheses concerning the unknown mean of a normal distribution. *Ann. Math. Stat.* 34:549–554.

Paulson, E. (1964). A sequential procedure for selecting the population with the largest mean from k normal populations. *Ann. Math. Stat.* 35:174–180.

Pavlov, I. V. (1988). A sequential procedure for testing many composite hypotheses. *Theo. Probab. Appl.* 33:138–142.

Robbins, H. (1970). Sequential estimation of an integer mean. In: *Scientists at Work, Festschrift in Honor of Herman Wold*. Uppsala: Almqvist and Wiksells.

Robbins, H., Sobel, M., Starr, N. (1968). A sequential procedure for selecting the largest of k means. *Ann. Math. Stat.* 39:88–92.

Romano, J. P., Wolf, M. (2005a). Exact and approximate stepdown methods for multiple hypothesis testing. *J. Am. Stat. Assoc.* 100(469):94–108.

Romano, J. P., Wolf, M. (2005b). Stepwise multiple testing as formalized data snooping. *Econometrica* 73(4):1237–1282.

Simons, G. (1967). Lower bounds for average sample number of sequential multihypothesis tests. *Ann. Math. Stat.* 38:1343–1364.

Sobel, M., Wald, A. (1949). A sequential decision procedure for choosing one of three hypotheses concerning the unknown mean of a normal distribution. *Ann. Math. Stat.* 20:502–522.

Tang, D.-I., Geller, N. L. (1999). Closed testing procedures for group sequential clinical trials with multiple endpoints. *Biometrics* 55:1188–1192.

Tang, D.-I., Geller, N. L., Pocock, S. J. (1993). On the design and analysis of randomized clinical trials with multiple endpoints. *Biometrics* 49:23–30.

Tang, D.-I., Lin, S. P. (1997). An approximate likelihood ratio test for comparing several treatments to a control. *J. Am. Stat. Assoc.* 92:1155–1162.