Exact and Approximate Stepdown Methods for Multiple Hypothesis Testing

Joseph P. ROMANO and Michael WOLF

Consider the problem of testing k hypotheses simultaneously. In this article we discuss finite- and large-sample theory of stepdown methods that provide control of the familywise error rate (FWE). To improve on the Bonferroni method or on Holm's stepdown method, Westfall and Young made effective use of resampling to construct stepdown methods that implicitly estimate the dependence structure of the test statistics. However, their methods depend on an assumption known as "subset pivotality." Our goal here is to construct general stepdown methods that do not require such an assumption. To accomplish this, we take a close look at what makes stepdown procedures work; a key component is a monotonicity requirement of critical values. By imposing monotonicity on estimated critical values (which is not an assumption on the model but rather is an assumption on the method), we show how to construct stepdown tests that can be applied in a stagewise fashion so that at most k tests need to be computed. Moreover, at each stage, an intersection test that controls the usual probability of a type 1 error is calculated, which allows us to draw on an enormous resampling literature as a general means of test construction. In addition, it is possible to carry out this method using the same set of resamples (or subsamples) for each of the intersection tests.

KEY WORDS: Bootstrap; Familywise error rate; Multiple testing; Permutation test; Randomization test; Stepdown procedure; Subsampling.

1. INTRODUCTION

The main point of this article is to show how computer-intensive methods can be used to construct asymptotically valid tests of multiple hypotheses under very weak conditions. As in the case of single testing, bootstrap and other resampling methods offer viable nonparametric alternatives to constructing tests that require normality or other parametric assumptions. The treatise by Westfall and Young (1993) takes good advantage of resampling to estimate the joint distributions of multiple test statistics to construct valid and more efficient multiple-testing methods. However, their methods rely heavily on the assumption of subset pivotality. Thus the main goal of the present article is to show how to construct valid stepdown methods that do not require this assumption, while remaining computationally feasible.

Consider the problem of testing hypotheses $H_1, \ldots, H_k$. Suppose that corresponding p-values $\hat{p}_1, \ldots, \hat{p}_k$ are available. A starting point for a general method that controls the familywise error rate (FWE) is the Bonferroni method, which rejects any H_j for which $\hat{p}_j \leq \alpha/k$. Holm (1979) improved this single-stage procedure by the following stepdown method: Order the p-values as

$$\hat{p}_{(1)} \leq \hat{p}_{(2)} \leq \cdots \leq \hat{p}_{(k)},$$

and let $H_{(1)}, \ldots, H_{(k)}$ denote the corresponding hypotheses. If $\hat{p}_{(1)} \leq \alpha/k$, then reject $H_{(1)}$; otherwise, accept all hypotheses and stop. If continuing, reject $H_{(2)}$ if $\hat{p}_{(2)} \leq \alpha/(k-1)$; otherwise, stop testing and accept all remaining hypotheses; and so on. Then hypotheses $H_{(1)}, \ldots, H_{(r)}$ are rejected if $\hat{p}_{(j)} \leq \alpha/(k-j+1)$ for $j = 1, \ldots, r$, and the remaining are accepted if $\hat{p}_{(r+1)} > \alpha/(k-r)$. This procedure holds under arbitrary dependence on the joint distribution of p-values. As was shown by Westfall and Young (1993), the Holm procedure can be improved by incorporating or estimating the dependence structure into the algorithm.

In Section 2 we discuss stepdown methods that control the FWE in finite samples. Such methods proceed stagewise by testing an intersection hypothesis at each stage. That is, like the Holm method, once a hypothesis is rejected, testing of the remaining hypotheses is accomplished as if the remaining hypotheses were a new family of joint hypotheses to be tested. Moreover, the decision to reject an hypothesis at a given stage depends only on the outcome of the intersection test for that stage.

But one cannot always achieve strong control in such a simple manner. By understanding the limitations of this approach in finite samples, we can then see why an asymptotic approach will be valid under fairly weak assumptions. It turns out that a simple monotonicity condition for theoretical critical values allows for some immediate results.

For any $K \subset \{1, \ldots, k\}$, let H_K denote the hypothesis that all H_j with $j \in K$ are true. The closure method of Marcus, Peritz, and Gabriel (1976) allows us to construct methods that control the FWE if we know how to test each intersection hypothesis H_K. However, in general, this might require the construction of $2^k - 1$ tests. The constructions studied here require only an order-k number of tests. In fact, the monotonicity assumptions that we invoke can be viewed as justification of an order-k stagewise application of closure. [In some cases, shortcuts for applying the closure method are known. For example, Westfall, Zaykin, and Young (2001) showed how to apply closure to Fisher combination tests with only k^2 evaluations.] A further advantage of our constructions is that they lead to consonant multiple-testing procedures in the sense of Hommel (1986); if the intersection hypothesis H_K is rejected, then necessarily at least one of the hypotheses H_j with $j \in K$ will be rejected. This property is appealing but does not always hold for the closure method of Marcus et al. (1976).

In general, we suppose that rejection of a test of H_j is based on large values of a test statistic $T_{n,j}$. (To be consistent with later

Joseph P. Romano is Professor, Department of Statistics, Stanford University, Stanford, CA 94305 (E-mail: *romano@stat.stanford.edu*). Michael Wolf is Assistant Professor, Department of Economics and Business, University of Pompeau Fabra, E-08005, Barcelona, Spain (E-mail: *michael.wolf@upf.edu*). This research was supported by National Science Foundation grant DMS-01-0392, Spanish Ministry of Science and Technology and FEDER grant BMF2003-03324, and the Barcelona Economics Program of CREA. The authors thank an associate editor for helpful comments that have led to an improved presentation of the article.

Journal of the American Statistical Association
March 2005, Vol. 100, No. 469, Theory and Methods
DOI 10.1198/016214504000000539

notation, we use n for asymptotic purposes and typically to refer to sample size.) Of course, if a p-value $\hat{p}_j$ is available for testing H_j, one possibility is to take $T_{n,j} = 1 - \hat{p}_j$. Then we restrict attention to tests that reject an intersection hypothesis H_K when $\max\{T_{n,j} : j \in K\}$ is large. In some problems where a monotonicity condition holds (distinct from the monotonicity assumption here), Lehmann, Romano, and Shaffer (2003) showed that such stepwise procedures are optimal in a maximin sense. In other situations, it may be better to consider other test statistics that combine the individual test statistics in a more powerful way. A related issue is one of balance; see Remark 8 in Section 4.2. At this time, our primary goal is to show how stepdown procedures can be constructed quite generally without having to assume subset pivotality, but while still controlling the FWE.

In Section 3 we show that if we estimate critical values that have a monotonicity property, then the basic problem of constructing a valid multiple-test procedure can be reduced to the problem of sequentially constructing critical values for (at most order k) single tests. This then allows us to directly apply what we know about tests based on permutation and randomization distributions. Similarly, we can apply bootstrap and subsampling methods as well, as we discuss in Section 4.

In Sections 5 and 6 we present two small simulation studies and an empirical application. We collect all proofs in an Appendix.

As noted previously, the closure method of Marcus et al. (1976) in principle allows us to reduce the problem of constructing a valid multiple-test procedure that controls the FWE to the problem of constructing a single test that controls the usual probability of a type 1 error; however, the number of tests that must be calculated increases exponentially with k. In general, if we wish to calculate a bootstrap test for each intersection hypothesis, then there is not only the computational issue of constructing a large number of tests, but also the question of an appropriate resampling mechanism that obeys the null hypothesis for each intersection hypothesis (unless there is a strong assumption, such as subset pivotality). Our methods are computationally feasible and avoid the need for a distinct resampling mechanism for each hypothesis. Thus this work is a sustained essay designed to produce computationally feasible general test constructions that control the FWE by effective reduction to the problem of construction of single tests that control the usual probability of a type 1 error. This then allows us to draw on an enormous resampling literature.

2. NONASYMPTOTIC RESULTS

Suppose that data $\mathbf{X}$ are generated from some unknown probability distribution P. In anticipation of asymptotic results, we may write $\mathbf{X} = \mathbf{X}^{(n)}$, where n typically refers to the sample size. A model assumes that P belongs to a certain family of probability distributions Ω, although we make no rigid requirements for Ω. Indeed, Ω may be a nonparametric model, a parametric model, or a semiparametric model.

Consider the problem of simultaneously testing a hypothesis H_j against H_j' for $j = 1, \ldots, k$. Of course, a hypothesis H_j can be viewed as a subset ω_j of Ω, in which case the hypothesis H_j is equivalent to $P \in \omega_j$ and H_j' is equivalent to $P \notin \omega_j$. For any subset $K \subset \{1, \ldots, k\}$, let $H_K = \bigcap_{j \in K} H_j$ be the hypothesis that $P \in \bigcap_{j \in K} \omega_j$.

Suppose that a test of the individual hypothesis H_j is based on a test statistic $T_{n,j}$, with large values indicating evidence against the H_j. For an individual hypothesis, numerous approaches to approximating a critical value exist, including those based on classical likelihood theory, bootstrap tests, Edgeworth expansions, permutation tests, and so on. The main problem addressed in the present work is to construct a procedure that controls the FWE. Recall that the FWE is the probability of rejecting at least one true null hypothesis. More specifically, if P is the true probability mechanism, then let $I = I(P) \subset \{1, \ldots, k\}$ denote the indices of the set of true hypotheses; that is, $j \in I$ if and only $P \in \omega_j$. The FWE is the probability under P that any H_j with $j \in I$ is rejected. To show its dependence on P, we may write $\mathrm{FWE} = \mathrm{FWE}_P$. We require that any procedure satisfy that the FWE be no bigger than α (at least asymptotically). Furthermore, this constraint must hold for all possible configurations of true and null hypotheses; that is, we demand strong control of the FWE. A procedure that controls the FWE only when all k null hypotheses are true is said to have weak control of the FWE. As noted by Dudoit, Shaffer, and Boldrick (2003), this distinction is often ignored.

For any subset K of $\{1, \ldots, k\}$, let $c_{n,K}(\alpha, P)$ denote an α-quantile of the distribution of $\max_{j \in K} T_{n,j}$ under P. Concretely,

$$c_{n,K}(\alpha, P) = \inf\left\{x : P\left\{\max_{j \in K} T_{n,j} \le x\right\} \ge \alpha\right\}. \tag{1}$$

For testing the intersection hypothesis H_K, we need only approximate a critical value for $P \in \bigcap_{j \in K} \omega_j$. Because there may be many such P, we define

$$c_{n,K}(1-\alpha) = \sup\left\{c_{n,K}(1-\alpha, P) : P \in \bigcap_{j \in K} \omega_j\right\}. \tag{2}$$

To define $c_{n,K}(\alpha, P)$, we implicitly assumed that $\bigcap_{j=1}^{k} \omega_j$ is not empty. At this point, we acknowledge that calculating these constants may be formidable in some problems (which is why we later turn to approximate or asymptotic methods).

Let

$$T_{n,r_1} \ge T_{n,r_2} \ge \cdots \ge T_{n,r_k} \tag{3}$$

denote the observed ordered test statistics, and let $H_{r_1}, H_{r_2}, \ldots, H_{r_k}$ be the corresponding hypotheses.

Stepdown procedures begin by testing the joint null hypothesis $H_{\{1,\ldots,k\}}$ that all hypotheses are true. This hypothesis is rejected if T_{n,r_1} is large. If it is not large, then accept all hypotheses; otherwise, reject the hypothesis corresponding to the largest test statistic. Once a hypothesis is rejected, remove it and test the remaining hypotheses by rejecting for large values of the maximum of the remaining test statistics, and so on. Thus at any step, one tests an intersection hypothesis, and an ideal situation would be to proceed at any step without regard to previous rejections, in the sense that once a hypothesis is rejected, the remaining hypotheses are treated as a new family, and testing for this new family proceeds independent of past decisions in such a way that rejecting one of the remaining hypotheses is based solely on the rejection of the next intersection test calculated. Because the Holm procedure (discussed later in Example 4) works in this way, one might hope that the intersection hypothesis can be generally tested at any step by treating only

those hypotheses that remain. Forgetting about whether or not such an approach generally yields strong control for the time being, we consider the following conceptual algorithm, which proceeds in stages by testing intersection hypotheses.

Algorithm 1 (Idealized stepdown method).

1. Let $K_1 = \{1, \ldots, k\}$. If $T_{n.r_1} \le c_{n.K_1}(1-\alpha)$, then accept all hypotheses and stop; otherwise, reject H_{r_1} and continue.
2. Let K_2 be the indices of the hypotheses not previously rejected. If $T_{n,r_2} \le c_{n,K_2}(1-\alpha)$, then accept all remaining hypotheses and stop; otherwise, reject H_{r_2} and continue.

⋮

j. Let K_j be the indices of the hypotheses not previously rejected. If $T_{n,r_j} \le c_{n,K_j}(1-\alpha)$, then accept all remaining hypotheses and stop; otherwise, reject H_{r_j} and continue.

⋮

k. If $T_{n,k} \le c_{n,K_k}(1-\alpha)$, then accept H_{r_k}; otherwise, reject H_{r_k}.

Algorithm 1 is an idealization for two reasons: (1) The critical values may be impossible to compute and (2) without restriction, there is no general reason why such a stepwise approach strongly controls the FWE. The determination of conditions where the algorithm leads to strong control will help us understand the limitations of a stepdown approach, as well as understand how such a general approach can work at least approximately in large samples. First, we present an example showing that some condition is required to exhibit strong control.

Example 1. Suppose that $T_{n,1}$ and $T_{n,2}$ are independent and normally distributed, with $T_{n,1} \sim \mathrm{N}(\theta_1, (1+\theta_2)^{2q})$ and $T_{n,2} \sim \mathrm{N}(\theta_2, (1+\theta_2)^{-2q})$, where $\theta_1 \ge 0$ and $\theta_2 \ge 0$. (The index n plays no role here, but we retain it for consistent notation.) Here q is a suitable positive constant, chosen to be large. Also, let $\Phi(\cdot)$ denote the standard normal cumulative distribution function. The hypothesis H_j specifies $\theta_j = 0$, whereas H_j' specifies $\theta_j > 0$. Therefore, the first step of Algorithm 1 is to reject the overall joint hypothesis $\theta_1 = \theta_2 = 0$ for large values of $\max(T_{n,1}, T_{n,2})$ when $T_{n,1}$ and $T_{n.2}$ are iid N(0, 1). Specifically, accept both hypotheses if

$$\max(T_{n,1}, T_{n,2}) \le c(1-\alpha) \equiv \Phi^{-1}(\sqrt{1-\alpha});$$

otherwise, reject the hypothesis corresponding to the larger $T_{n,j}$. Such a procedure exhibits weak control but not strong control. For example, the probability of rejecting the H_1 at the first step when $\theta_1 = 0$ and $\theta_2 = c(1-\alpha)/2$ satisfies

$$P_{0,\theta_2}\{T_{n.1} > c(1-\alpha),\ T_{n.1} > T_{n.2}\} \to 1/2$$

as $q \to \infty$. So if $\alpha < 1/2$, for some sufficiently large but fixed q, then the probability of incorrectly declaring H_1 to be false is greater than α. Incidentally, this also provides an example of a single-step procedure that exhibits weak control but not strong control. (Single-step procedures are those in which hypotheses are rejected on the basis of a single critical value; see Westfall and Young 1993.)

Therefore, to prove strong control, some condition is required. Consider the following monotonicity assumption: For $I \subset K$,

$$c_{n,K}(1-\alpha) \ge c_{n,I}(1-\alpha). \tag{4}$$

The condition (4) can be expected to hold in many situations, because the left side is based on computing the $1-\alpha$ quantile of the maximum of $|K|$ variables, whereas the right side is based on the maximum of $|I| \le |K|$ variables (although one must be careful and realize that the quantiles are computed under possibly different P, which is why some conditioning is required).

Theorem 1. Let P denote the true distribution generating the data. Assume that $\bigcap_{j=1}^{k} \omega_j$ is not empty.

(a) Assume that for any K containing $I(P)$,

$$c_{n,K}(1-\alpha) \ge c_{n,I(P)}(1-\alpha). \tag{5}$$

Then the probability that Algorithm 1 rejects any $j \in I(P)$ is $\le \alpha$; that is, $\mathrm{FWE}_P \le \alpha$.

(b) Strong control persists if, in Algorithm 1, the critical constants $c_{n,K_j}(1-\alpha)$ are replaced by $d_{n,K_j}(1-\alpha)$, which satisfy

$$d_{n.K_j}(1-\alpha) \ge c_{n,K_j}(1-\alpha). \tag{6}$$

(c) Moreover, condition (5) may be removed if the $d_{n,K_j}(1-\alpha)$ satisfy

$$d_{n,K}(1-\alpha) \ge d_{n,I(P)}(1-\alpha) \tag{7}$$

for any $K \supset I(P)$.

Remark 1. Under weak assumptions, we can show the sup over P of the probability (under P) that Algorithm 1 rejects any $j \in I(P)$ is equal to α. It then follows that the critical values cannot be made smaller in a hope of increasing the ability to detect false hypotheses without violating the strong control of the FWE. (However, this does not negate the possibility of smaller random critical values, as long as they are not smaller with probability 1.)

Example 2 (Assumption of subset pivotality). Assumptions stronger than (5) have been used. Suppose, for example, that for every subset $K \subset \{1, \ldots, k\}$ there exists a distribution P_K that satisfies

$$c_{n,K}(1-\alpha, P) \le c_{n,K}(1-\alpha, P_K) \tag{8}$$

for all P such that $I(P) \supset K$. Such a P_K may be viewed as being least favorable among distributions P such that $P \in \bigcap_{j \in K} \omega_j$. (For example, if H_j corresponds to a parameter $\theta_j \le 0$, then intuition suggests that a least-favorable configuration should correspond to $\theta_j = 0$.)

In addition, assume the subset pivotality condition of Westfall and Young (1993); that is, assume that there exists a P_0 with $I(P_0) = \{1, \ldots, k\}$ such that the joint distribution of $\{T_{n,j} : j \in I(P_K)\}$ under P_K is the same as the distribution of $\{T_{n,j} : j \in I(P_K)\}$ under P_0. This condition says the (joint) distribution of the test statistics used for testing the hypotheses H_j, $j \in I(P_K)$, that is unaffected by the truth or falsehood of the remaining hypotheses (and therefore we assume that all hypotheses are true

by calculating the distribution of the maximum under P_0). It follows that in step j of Algorithm 1,

$$c_{n,K_j}(1-\alpha) = c_{n,K_j}\big(1-\alpha, P_{K_j}\big) = c_{n,K_j}(1-\alpha, P_0) = c_{n,K_j}(1-\alpha). \quad (9)$$

The outer equalities in (9) follow by the assumption (8), and the middle equality follows by the subset pivotality condition. Therefore, in Algorithm 1, we can replace $c_{n,K_j}(1-\alpha)$ by $c_{n,K_j}(1-\alpha, P_0)$, which in principle is known because it is the $1-\alpha$ quantile of the distribution of $\max(T_{n,j}: j \in K_j)$ under P_0, and P_0 is some fixed (least favorable) distribution. At the very least, this quantile may be simulated.

The asymptotic behavior of stepwise procedures was considered by Finner and Roters (1998), who recognized the importance of monotonicity for the validity of stepwise procedures. But they also supposed the existence of a single least-favorable P_0 for all configurations of true hypotheses, which then guarantees monotonicity of critical values for stepdown procedures. As seen previously, such assumptions do not hold generally.

Example 3. To exhibit an example in which condition (5) holds but subset pivotality does not, suppose that $T_{n,1}$ and $T_{n,2}$ are independent and normally distributed, with $T_{n,1} \sim \mathrm{N}(\theta_1, 1/(1+\theta_2^2))$ and $T_{n,2} \sim \mathrm{N}(\theta_2, 1/(1+\theta_1^2))$. The hypothesis H_j specifies $\theta_j = 0$, whereas the alternative H_j' specifies $\theta_j > 0$. Then it is easy to check that, with $K_1 = \{1, 2\}$,

$$c_{n,K_1}(1-\alpha) = \Phi^{-1}(\sqrt{1-\alpha}\,) > \Phi^{-1}(1-\alpha) = c_{n,\{j\}}(1-\alpha).$$

Therefore, (5) holds, but subset pivotality fails.

Example 4 (The Holm procedure). Suppose that $-T_{n,j} \equiv \hat{p}_{n,j}$ is a p-value for testing H_j; that is, assume that the distribution of $\hat{p}_{n,j}$ is uniform on $(0, 1)$ when H_j is true. Note that this assumption is much weaker than subset pivotality (if $k > 1$), because we are making an assumption only about the one-dimensional marginal distribution of the p-value statistic. Furthermore, we may assume the weaker condition

$$P\{\hat{p}_{n,j} \le x\} \le x$$

for any $x \in (0, 1)$ and any $P \in \omega_j$. If $I(P) \supset K$, then the usual argument using the Bonferroni inequality yields

$$c_{n,K}(1-\alpha, P) \le -\alpha/|K|,$$

which is independent of P, and so

$$c_{n,K}(1-\alpha) \le -\alpha/|K|. \quad (10)$$

It is easy to construct joint distributions for which this is attained, and so we have equality here if the family Ω is so large that it includes all possible joint distributions for the p-values. In such a case, we have equality in (10), and so the condition (5) is satisfied. Of course, even if the model is not so large, this procedure has strong control. Simply, let $d_{n,K}(1-\alpha) = -\alpha/|K|$, and strong control follows by Theorem 1, part (c).

Part (c) of Theorem 1 points toward a more general method that has strong control even when (5) is violated and can be much less conservative than the Holm procedure.

Corollary 1. Assume that $\bigcap_{j=1}^{k} \omega_j$ is not empty. Let

$$c^*_{n,K_j}(1-\alpha) = \max\{c_{n,K}(1-\alpha): K \subset K_j\}. \quad (11)$$

Then, if we replace $c_{n,K_j}(1-\alpha)$ by $c^*_{n,K_j}(1-\alpha)$ in Algorithm 1, strong control holds.

Corollary 1 is simply the closure principle of Marcus et al. (1976) (also see Hommel 1986 and thm. 4.1 of Hochberg and Tamhane 1987). Thus, to have a valid stepdown procedure, one not only must consider the critical value $c_{n,K}(1-\alpha)$ when testing an intersection hypothesis H_K, but also must compute all $c_{n,I}(1-\alpha)$ for $I \subset K$.

Finally, we can remove the assumption that $\bigcap_{j=1}^{k} \omega_j$ is not empty, as follows.

Theorem 2. Let

$$\tilde{c}_{n,K_j}(1-\alpha) = \max\left\{c_{n,K}(1-\alpha): K \subset K_j \text{ and } \bigcap_{j\in K} \omega_j \ne \emptyset\right\}. \quad (12)$$

(a) If $c_{n,K_j}(1-\alpha)$ are replaced by $\tilde{c}_{n,K_j}(1-\alpha)$ in Algorithm 1, then strong control holds.

(b) Strong control persists if, in Algorithm 1, the critical constants $c_{n,K_j}(1-\alpha)$ are replaced by $\tilde{d}_{n,K_j}(1-\alpha)$, which satisfy

$$\tilde{d}_{n,K_j}(1-\alpha) \ge \tilde{c}_{n,K_j}(1-\alpha). \quad (13)$$

Theorem 2 shows that the Holm method applies with no assumptions; that is, the assumption that all hypotheses can be true need not hold.

3. RANDOM CRITICAL VALUES AND RANDOMIZATION TESTS

3.1 Preliminaries and a Basic Inequality

In general, the critical values used in Algorithm 1 are the smallest constants possible without violating the FWE. As a simple example, suppose that X_j, $j = 1, \ldots, k$, are independent $\mathrm{N}(\theta_j, 1)$, with the θ_j varying freely. The null hypothesis H_j specifies $\theta_j \le 0$. Then

$$c_{n,K}(1-\alpha) = \Phi^{-1}\big[(1-\alpha)^{(1/|K|)}\big].$$

Suppose that c is a constant and that $c < c_{n,K}(1-\alpha)$ for some subset K. As $\theta_j \to \infty$ for $j \notin K$ and $\theta_j = 0$ for $j \in K$, the probability of a type 1 error tends to

$$P\left\{\max_{j\in K} X_j > c\right\} > P\left\{\max_{j\in K} X_j > c_{n,K}(1-\alpha)\right\} = \alpha.$$

Of course, if the θ_j are bounded, then the argument fails, but typically such assumptions are not made.

However, the foregoing applies only to nonrandom critical values and leaves open the possibility that critical values can be estimated and thus can be random. That is, if we replace $c_{n,K}(1-\alpha)$ by some estimate $\hat{c}_{n,K}(1-\alpha)$, then it sometimes can be smaller than $c_{n,K}(1-\alpha)$ as long as it is not with probability 1. Of course, this is the typical case in which critical values must be estimated, such as by the bootstrap in the next section. In this section we focus on the use of permutation and

randomization tests that replace the idealized critical values by estimated ones, while still retaining finite-sample control of the FWE.

One simple way to deal with permutation and randomization tests is to define critical values conditional on an appropriate σ-field; then the monotonicity assumptions of the previous section will turn into monotonicity assumptions for the conditional critical values. (For example, in the context of comparing two samples, everything would be conditional on the values of the combined sample, and this would lead directly to permutation tests.)

For the sake of increased generality, we instead proceed as follows. Suppose that the $c_{n,K}(1-\alpha)$ in Algorithm 1 are replaced by estimates $\hat{c}_{n,K}(1-\alpha)$. These could be obtained by a permutation test if it applies, but for the moment their construction is left unspecified. However, we assume two things. First, we replace the monotonicity assumption (5) by monotonicity of the estimated critical values; that is, for any $K \supset I(P)$,

$$\hat{c}_{n,K}(1-\alpha) \geq \hat{c}_{n,I(P)}(1-\alpha). \tag{14}$$

We then also require that if $\hat{c}_{n,K}(1-\alpha)$ is used to test the intersection hypothesis H_K, then it is level α when $K = I(P)$; that is,

$$P\{\max(T_{n,j} : j \in I(P)) > \hat{c}_{n,I(P)}(1-\alpha)\} \leq \alpha. \tag{15}$$

We show the basic inequality that the FWE_P is bounded above by the left side of (15). This will then show that, if we can construct monotone critical values such that each intersection test is level α, then the stepdown procedure controls the FWE. Thus construction of a stepdown procedure is effectively reduced to construction of single tests, as long as the monotonicity assumption holds. (Also note the monotonicity assumption for the critical values, which is something that we can essentially enforce because they depend only on the data, can hold even if the corresponding nonrandom ones are not monotone.) Note that here (and in the rest of the article), we no longer need to assume $\bigcap_{j=1}^{k} \omega_j$ is not empty.

Theorem 3. Let P denote the true distribution generating the data. Consider Algorithm 1 with $c_{n,K}(1-\alpha)$ replaced by estimates $\hat{c}_{n,K}(1-\alpha)$ satisfying (14).

(a) Then

$$\mathrm{FWE}_P \leq P\{\max(T_{n,j} : j \in I(P)) > \hat{c}_{n,I(P)}(1-\alpha)\}. \tag{16}$$

(b) Therefore, if the critical values also satisfy (15), then $\mathrm{FWE}_P \leq \alpha$.

3.2 Permutation and Randomization Tests

Before applying Theorem 3, we first review a general construction of a randomization test in the context of a single test. Our setup is framed in terms of a population model, but similar results are possible in terms of a randomization model (as in sec. 3.1.7 of Westfall and Young 1993).

Based on data $\mathbf{X}$ taking values in a sample space $\mathcal{X}$, it is desired to test the null hypothesis H that the underlying probability law P generating $\mathbf{X}$ belongs to a certain family ω of distributions. Let $\mathbf{G}$ be a finite group of transformations g of $\mathcal{X}$ onto itself. The following assumption, which we call the *randomization hypothesis*, allows for a general test construction.

The Randomization Hypothesis. The null hypothesis implies that the distribution of $\mathbf{X}$ is invariant under the transformations in $\mathbf{G}$; that is, for every g in $\mathbf{G}$, $g\mathbf{X}$ and $\mathbf{X}$ have the same distribution whenever $\mathbf{X}$ has distribution P in ω.

As an example, consider testing the equality of distributions based on two independent samples $(\mathbf{Y}_1, \ldots, \mathbf{Y}_m)$ and $(\mathbf{Z}_1, \ldots, \mathbf{Z}_n)$. Under the null hypothesis that the samples are generated from the same probability law, the observations can be permuted or assigned at random to either of the two groups, and the distribution of the permuted samples is the same as the distribution of the original samples. In this example, and more generally when the randomization hypothesis holds, the following construction of a randomization test applies.

Let $T(\mathbf{X})$ be any real-valued test statistic for testing H. Suppose that the group $\mathbf{G}$ has M elements. Given $\mathbf{X} = \mathbf{x}$, let

$$T^{(1)}(\mathbf{x}) \leq T^{(2)}(\mathbf{x}) \leq \cdots \leq T^{(M)}(\mathbf{x})$$

be the values of $T(g\mathbf{x})$ as g varies in $\mathbf{G}$, ordered from smallest to largest. Fix a nominal level α, $0 < \alpha < 1$, and let m be defined by

$$m = M - [M\alpha], \tag{17}$$

where $[M\alpha]$ denotes the largest integer less than or equal to $M\alpha$. Let $M^+(\mathbf{x})$ and $M^0(\mathbf{x})$ be the number of values $T^{(j)}(\mathbf{x})$ $(j = 1, \ldots, M)$ that are greater than $T^{(m)}(\mathbf{x})$ and equal to $T^{(m)}(\mathbf{x})$. Set

$$a(\mathbf{x}) = \frac{M\alpha - M^+(\mathbf{x})}{M^0(\mathbf{x})}.$$

Define the randomization test function $\phi(\mathbf{X})$ to be equal to 1, $a(\mathbf{X})$, or 0 according to whether $T(\mathbf{X}) > T^{(m)}(\mathbf{X})$, $T(\mathbf{X}) = T^{(m)}(\mathbf{X})$, or $T(\mathbf{X}) < T^{(m)}(\mathbf{X})$.

Under the randomization hypothesis, Hoeffding (1952) showed that this construction produces a test that is exact level α, and this result is true for *any* choice of test statistic T. Note that this test is possibly a randomized test if $M\alpha$ is not an integer of there are ties in the ordered values. Alternatively, if one prefers not to randomize, then the slightly conservative but *nonrandomized* test that rejects if $T(\mathbf{X}) > T^m(\mathbf{X})$ is level α.

For any $\mathbf{x} \in \mathcal{X}$, let $\mathbf{G}^{\mathbf{x}}$ denote the $\mathbf{G}$-orbit of $\mathbf{x}$, that is,

$$\mathbf{G}^{\mathbf{x}} = \{g\mathbf{x} : g \in \mathbf{G}\}.$$

These orbits partition the sample space. Then, under the randomization hypothesis, it can be shown that the conditional distribution of $\mathbf{X}$ given $\mathbf{X} \in \mathbf{G}^{\mathbf{x}}$ is uniform on $\mathbf{G}^{\mathbf{x}}$.

In general, a p-value, $\hat{p}$, of a randomization test can be defined by

$$\hat{p} = \frac{1}{M} \sum_{g} I\{T(g\mathbf{X}) \geq T(\mathbf{X})\}. \tag{18}$$

It is easily shown that under the null hypothesis, $\hat{p}$ satisfies

$$P\{\hat{p} \leq u\} \leq u \quad \text{for all } 0 \leq u \leq 1. \tag{19}$$

Therefore, the *nonrandomized* test that rejects when $\hat{p} \leq \alpha$ is level α.

Because $\mathbf{G}$ may be large, we may resort to a stochastic approximation to construct the randomization test by, for example, by randomly sampling transformations g from $\mathbf{G}$ with or

without replacement. In the former case, for example, suppose that $g_1, \ldots, g_{B-1}$ are iid and uniformly distributed on $\mathbf{G}$. Let

$$\tilde{p} = \frac{1}{B}\left[1 + \sum_{i=1}^{B-1} I\{T(g_i\mathbf{X}) \geq T(\mathbf{X})\}\right]. \tag{20}$$

Then it can be shown that under the randomization hypothesis,

$$P\{\tilde{p} \leq u\} \leq u \quad \text{for all } 0 \leq u \leq 1, \tag{21}$$

where this probability reflects variation in both $\mathbf{X}$ and the sampling of the g_i. Note that (21) holds for any B, and so the test that rejects when $\tilde{p} \leq \alpha$ is level α even when a stochastic approximation is used. Of course, the larger the value of B, the closer $\hat{p}$ and $\tilde{p}$ are to each other; in fact, $\hat{p} - \tilde{p} \to 0$ in probability as $B \to \infty$. The argument for (20) is based on the following simple fact.

Lemma 1. Suppose that $\mathbf{Y}_1, \ldots, \mathbf{Y}_B$ are exchangeable real-valued random variables; that is, their joint distribution is invariant under permutations. Let $\tilde{q}$ be defined by

$$\tilde{q} = \frac{1}{B}\left[1 + \sum_{i=1}^{B-1} I\{\mathbf{Y}_i \geq \mathbf{Y}_B\}\right].$$

Then $P\{\tilde{q} \leq u\} \leq u$ for all $0 \leq u \leq 1$.

We now return to the multiple testing problem. Assume that $\mathbf{G}_K$ is a group of transformations for which the randomization hypothesis holds for H_K. Then we can apply the foregoing construction to test the single-intersection hypothesis H_K based on the test statistic

$$T_{n,K} = \max(T_{n,j} : j \in K), \tag{22}$$

and reject H_K when

$$T_{n,K}(\mathbf{X}) > T_{n,K}^{(|\mathbf{G}_K| - [|\mathbf{G}_K|\alpha])}(\mathbf{X}).$$

If we further specialize to the case where $\mathbf{G}_K = \mathbf{G}$, so that the same $\mathbf{G}$ applies to all intersection hypotheses, then we can verify the monotonicity assumption for the critical values. Set $m_\alpha = |\mathbf{G}| - [|\mathbf{G}|\alpha]$. Then, for any $g \in \mathbf{G}$ and $I \subset K$,

$$\max\bigl(T_{n,j}(g\mathbf{X}) : j \in K\bigr) \geq \max\bigl(T_{n,j}(g\mathbf{X}) : j \in I\bigr), \tag{23}$$

and so as g varies, the m_αth-largest value of the left side of (23) is at least as large as the m_αth-largest value of the right side.

Consequently, the critical values

$$\hat{c}_{n,K}(1-\alpha) = T_{n,K}^{(m_\alpha)} \tag{24}$$

satisfy the monotonicity requirement of Theorem 3. Moreover, by the general randomization construction of a single test, the test that rejects H_K when $T_K \geq T_{n,K}^{(m_\alpha)}$ is level α. Therefore, the following result is true.

Corollary 2. Suppose that the randomization hypothesis holds for a group $\mathbf{G}$ when testing any intersection hypothesis H_K. Then the stepdown method with critical values given by (24) controls the FWE.

Equivalently, in analogy with (18), we can compute p-values for testing H_K via

$$\hat{p}_{n,K} = \frac{1}{M}\sum_{g} I\{T_{n,K}(g\mathbf{X}) \geq T_{n,K}(\mathbf{X})\}, \tag{25}$$

and at stage j where we are testing an intersection hypothesis, say H_K, reject if $\hat{p}_{n,K} \leq \alpha$. Alternatively, we can approximate these p-values and still retain the level of the test. In analogy with (20), randomly sample $g_1, \ldots, g_{B-1}$ from $\mathbf{G}$ and let

$$\tilde{p}_{n,K} = \frac{1}{B}\left[1 + \sum_{i=1}^{B-1} I\{T_{n,K}(g_i\mathbf{X}) \geq T_{n,K}(\mathbf{X})\}\right]. \tag{26}$$

By an almost identical argument, we have the following result.

Corollary 3. Suppose that the randomization hypothesis holds for a group $\mathbf{G}$ when testing any intersection hypothesis H_K. Consider the stepdown method that rejects K_j at stage j if $\tilde{p}_{n,K_j} \leq \alpha$. Then $\mathrm{FWE}_P \leq \alpha$.

Remark 2. In the foregoing corollaries, we have worked with the randomization construction using nonrandomized tests. A similar result would hold if we would permit randomization.

Example 5 (Two-sample problem with k variables). Suppose that $\mathbf{Y}_1, \ldots, \mathbf{Y}_{n_\mathbf{Y}}$ is a sample of $n_\mathbf{Y}$ independent observations from a probability distribution $P_\mathbf{Y}$ and that $\mathbf{Z}_1, \ldots, \mathbf{Z}_{n_\mathbf{Z}}$ is a sample of $n_\mathbf{Z}$ observations from $P_\mathbf{Z}$. Here $P_\mathbf{Y}$ and $P_\mathbf{Z}$ are probability distributions on $\mathbb{R}^k$, with the jth components denoted by $P_{\mathbf{Y},j}$ and $P_{\mathbf{Z},j}$. The hypothesis H_j asserts that $P_{\mathbf{Y},j} = P_{\mathbf{Z},j}$, and we wish to test these k hypotheses based on $\mathbf{X} = (\mathbf{Y}_1, \ldots, \mathbf{Y}_{n_\mathbf{Y}}, \mathbf{Z}_1, \ldots, \mathbf{Z}_{n_\mathbf{Z}})$. Also, let $Y_{i,j}$ denote the jth component of $\mathbf{Y}_i$ and let $Z_{i,j}$ denote the jth component of $\mathbf{Z}_i$. Following Troendle (1995), we assume a semiparametric model. In particular, assume that $P_\mathbf{Y}$ and $P_\mathbf{Z}$ are governed by a family of probability distributions $Q_{\boldsymbol{\theta}}$ indexed by $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k) \in \mathbb{R}^k$ (and assumed identifiable), so that $P_\mathbf{Y}$ has law $Q(\boldsymbol{\theta}_\mathbf{Y})$ and $P_\mathbf{Z}$ has law $Q(\boldsymbol{\theta}_\mathbf{Z})$. For concreteness, we may think of $\boldsymbol{\theta}$ as being the mean vector, although this assumption is not necessary. Now H_j can be viewed as testing $\theta_{\mathbf{Y},j} = \theta_{\mathbf{Z},j}$. Note that the randomization construction does not need to assume knowledge of the form of Q (just as a single two-sample permutation test in a shift model does not need to know the form of the underlying distribution under the null hypothesis).

Let $n = n_\mathbf{Y} + n_\mathbf{Z}$, and for $\mathbf{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n$, let $g\mathbf{x} \in \mathbb{R}^N$ be defined by $(x_{\pi(1)}, \ldots, x_{\pi(n)})$, where $(\pi(1), \ldots, \pi(n))$ is a permutation of $(1, 2, \ldots, n)$. Let $\mathbf{G}$ be the collection of all such g so that $M = n!$. Under the hypothesis $P_\mathbf{Y} = P_\mathbf{Z}$, $g\mathbf{X}$ and $\mathbf{X}$ have the same distribution for any g in $\mathbf{G}$.

Unfortunately, this $\mathbf{G}$ does not apply to any subset of the hypotheses, because $g\mathbf{X}$ and $\mathbf{X}$ need not have the same distribution if only a subcollection of the hypotheses are true. However, we just need a slight generalization to cover the example. Suppose that the test statistic $T_{n,j}$ used to test H_j depends only on the jth components of the observations, namely $Y_{i,j}$, $i = 1, \ldots, n_\mathbf{Y}$ and $Z_{i,j}$, $i = 1, \ldots, n_\mathbf{Z}$; this is a weak assumption indeed. In fact, let $\mathbf{X}_K$ be the dataset consisting of the components $Y_{i,j}$ and $Z_{i,j}$ as j varies only in K. The simple but important point here is that for this reduced dataset, the randomization hypothesis holds. Specifically, under the null hypothesis

$\theta_{\mathbf{Y},j} = \theta_{\mathbf{Z},j}$ for $j \in K$, $\mathbf{X}_K$ and $g\mathbf{X}_K$ have the same distribution (although $\mathbf{X}$ and $g\mathbf{X}$ need not). Also, for any $g \in \mathbf{G}$, $T_{n,j}(g\mathbf{X})$ and $T_{n,j}(\mathbf{X})$ have the same distribution under H_j, and similarly for any $K \subset \{1, \ldots, k\}$, $T_{n,K}(g\mathbf{X})$ and $T_{n,K}(\mathbf{X})$ have the same distribution under H_K.

Then, because the same $\mathbf{G}$ applies in this manner for all K, the critical values from the randomization test are monotone, just as in (23). Moreover, each intersection hypothesis can be tested by an exact level-α randomization test (because inference for H_K is based only on $\mathbf{X}_K$). Therefore, essentially the same argument leading to Corollaries 2 and 3 applies. In particular, even if we need to resort to approximate randomization tests at each stage, as long as we sample the same set of g_i from $\mathbf{G}$, the resulting procedure retains its finite-sample property of controlling the FWE. In contrast, Troendle (1995) concluded asymptotic control.

Remark 3. It is interesting to study the behavior of randomization procedures if the model is such that the randomization hypothesis does not hold. For example, in Example 5, suppose that we are interested just in testing the hypothesis H_j' that the mean of $P_{\mathbf{Y},j}$ is the mean of $P_{\mathbf{Z},j}$ (assumed to exist). Then the randomization test construction of this section fails, because the randomization hypothesis need not hold. However, because the randomization procedure has monotone critical values (because this is a property only of how the data are used), Theorem 3(a) applies. Therefore, one can again reduce the problem of studying control of the FWE to that of controlling the level of a single-intersection hypothesis. But the problem of controlling the level of a single test when the randomization hypothesis fails was studied by Romano (1990), and so similar methods can be used here, with the hope of at least proving asymptotic control. Alternatively, the more general resampling approaches of Section 4 can be used; the comparison of randomization and bootstrap tests was studied by Romano (1989), who showed they are often quite close, at least when the randomization hypothesis holds.

Example 6 (Problem of multiple treatments). Consider the one-way ANOVA model. We are given $k+1$ independent samples, with the jth sample having n_j iid observations $\mathbf{X}_{i,j}$, $i = 1, \ldots, n_j$. Suppose that $\mathbf{X}_{i,j}$ has distribution P_j. The problem is to test the hypotheses of k treatments with a control; that is, $H_j : P_j = P_{k+1}$. (Alternatively, we can test all pairs of distributions, but the issues are much the same, so we illustrate them with the slightly easier setup.) Under the joint null hypothesis, we can randomly assign all $n = \sum_j n_j$ observations to any of the groups; that is, the group $\mathbf{G}$ consists of all permutations of the data. However, if only a subset of the hypotheses are true, then this group is not valid. A simple remedy is to permute only within subsets; that is, to test any subset hypothesis H_K, consider only those permutations that permute observations within the sample $\mathbf{X}_{i,k+1}$ and the samples $\mathbf{X}_{i,j}$ with $j \in K$. Therefore, we compute a critical value by $\hat{c}_{n,K}(1-\alpha)$ by the randomization test with the group $\mathbf{G}_K$ of permutations within samples $j \in K$ and $j = k+1$. Unfortunately, this does not lead to monotonicity of critical values, and the previous results do not apply. But there is an analog of Corollary 1, if we are willing to compute critical values for all subset hypotheses; that is, replace $\hat{c}_{n,K_j}(1-\alpha)$ by

$$\hat{c}^*_{n,K_j}(1-\alpha) = \max\{\hat{c}_{n,K}(1-\alpha) : K \subset K_j\}.$$

But this approach can be computationally prohibitive. Such issues have been raised by Petrondas and Gabriel (1983) (although they did not frame the problem in terms of a monotonicity requirement). However, we shortly see that the lack of monotonicity of critical values is only a finite-sample concern; see Example 8.

4. ASYMPTOTIC RESULTS

The main goal of this section is to construct asymptotically valid stepdown procedures that hold under very weak assumptions, even when the monotonicity condition of Theorem 1 fails. The assumptions are identical to the weakest assumptions available for the construction of asymptotically valid tests of a single hypothesis, which are used in many resampling schemes, and so we cannot expect to improve them without improving the now well-developed theory of resampling methods for testing a single hypothesis.

Of course, Corollary 1 reminds us that it may be possible to construct a test that controls the FWE if we are willing and able to compute critical values for all possible $2^k - 1$ nontrivial intersection hypotheses. If each such test were computed by a bootstrap or resampling method, then the number of computations could get quite large for even moderate k. We not only provide weak conditions, but also consider a method that requires only *one* set of bootstrap resamples, as well as a method based on *one* set of subsamples.

To accomplish this without having to invoke an assumption like subset pivotality, we consider resampling schemes that do *not* obey the constraints of the null hypothesis. Schemes that do obey the constraints of the null hypothesis, as discussed by Beran (1986) and Romano (1988), are based on the idea that the critical value should be obtained under the null hypothesis, and so the resampling scheme should reflect the constraints of the null hypothesis. This idea was even advocated as a principle by Hall and Wilson (1991), and it was enforced by Westfall and Young (1993). Although appealing, it is by no means the only approach toward inference in hypothesis testing. Indeed, the well-known explicit duality between tests and confidence intervals means that if we can construct good or valid confidence intervals, then we can construct good or valid tests, and conversely. But resampling the empirical distribution to construct a confidence interval for a single parameter can produce very desirable intervals, which would then translate into desirable tests. The same holds for simultaneous confidence sets and multiple tests.

That is not to say that the approach of obeying the null constraints is less appealing. It is, however, often more difficult to apply, and it is unlikely that one resampling scheme obeying the constraints of all hypotheses would work in general in the multiple-testing framework. An alternative approach would be to resample from a different distribution at each step, obeying the constraints of the null hypotheses imposed at each step. This approach would probably succeed in a fair amount of generality, but even so, two problems would remain. First, it may be difficult to determine the appropriate resampling scheme for testing each subset hypothesis. Second, even if we knew how to resample at each stage, increased computation is involved. Our approach avoids these complications.

Before embarking on the general theory, we present a motivating example to fix ideas.

Example 7 (Testing correlations). Suppose that $\mathbf{X}_1, \ldots, \mathbf{X}_n$ are iid random vectors in $\mathbb{R}^s$, so that $\mathbf{X}_i = (X_{i,1}, \ldots, X_{i,s})$. Assume that $E|X_{i,j}|^2 < \infty$ and $\mathrm{Var}(X_{i,j}) > 0$, so that the correlation between $X_{1,i}$ and $X_{1,j}$, namely $\rho_{i,j}$, is well defined. Let $H_{i,j}$ denote the hypothesis that $\rho_{i,j} = 0$, so that the multiple-testing problem consists in testing all $k = \binom{s}{2}$ pairwise correlations. Also let $T_{n,i,j}$ denote the ordinary sample correlation between variables i and j. (Note that we are indexing hypotheses and test statistics by two indices, i and j.) As noted by Westfall and Young (1993), example 2.2, p. (43), subset pivotality fails here. For example, using results of Aitken (1969, 1971), if $s = 3$, $H_{1,2}$ and $H_{1,3}$ are true but $H_{2,3}$ is false, then the joint limiting distribution of $n^{1/2}(T_{n,1,2}, T_{n,1,3})$ is bivariate normal with mean 0, variance 1, and correlation $\rho_{2,3}$. As Westfall and Young (1993) acknowledged, their methods fail to address this problem (even asymptotically).

4.1 General Results

We now develop some asymptotic theory. For any $K \subset \{1, \ldots, k\}$, let $G_{n,K}(P)$ be the joint distribution of $T_{n,j}, j \in K$ under P, with corresponding joint cdf $G_{n,K}(\mathbf{x}, P)$, $\mathbf{x} \in \mathbb{R}^{|K|}$. Also, let $H_{n,K}(P)$ denote the distribution of $\max\{T_{n,j} : j \in K\}$ under P. As in the previous section, $c_{n,K}(1-\alpha, P)$ denotes a $1-\alpha$ quantile of $H_{n,K}(P)$. Also, the symbols $\xrightarrow{L}$ and $\xrightarrow{P}$ denote convergence in law (or distribution) and convergence in probability.

Typically, the asymptotic behavior of $G_{n,I(P)}(P)$ is governed by one of the following two possibilities: It either it has a nondegenerate limiting distribution or converges weakly to a nondegenerate constant vector (possibly with some components $-\infty$). Actually, this has nothing to do with the fact that we are studying joint distributions of multiple test statistics. For example, suppose that we are testing whether a population mean $\mu(P)$ is ≤ 0 versus >0 based on an iid sample $X_1, \ldots, X_n$ from P, assumed to have a finite nonzero variance $\sigma^2(P)$. Consider the test statistic $T_n = n^{-1/2} \sum_i X_i$. If $\mu(P) = 0$, then $T_n \xrightarrow{L} \mathrm{N}(0, \sigma^2(P))$. On the other hand, if $\mu(P) < 0$, then, T_n converges in probability to $-\infty$. Alternatively, if the test statistic is $T_n' = \max(0, T_n)$, then, if $\mu(P) = 0$, T_n' converges in distribution to $\max(0, \sigma(P)Z)$, where $Z \sim \mathrm{N}(0, 1)$. But under $\mu(P) < 0$, T_n' converges in probability to 0. Note that the two cases exhaust all possibilities under the null hypothesis. On the other hand, for the two-sided problem of testing $\mu(P) = 0$ versus $\mu(P) \neq 0$ based on $|n^{-1/2} \sum_i X_i|$, a nondegenerate limit law exists under the null hypothesis, and this exhausts all possibilities under the null hypothesis (under the assumption of a finite positive variance).

Formally, we distinguish between the following assumptions, which are imposed only when $K = I(P)$ is the set of true hypotheses.

Assumption A1. Under P, the joint distribution of the test statistics $T_{n,j}, j \in I(P)$, has a limiting distribution, that is,

$$G_{n,I(P)}(P) \xrightarrow{L} G_{I(P)}(P). \tag{27}$$

This implies that under P, $\max\{T_{n,j} : j \in I(P)\}$ has a limiting distribution, say $H_{I(P)}(P)$, with limiting cdf $H_{I(P)}(x, P)$. We further assume that

$$H_{I(P)}(x, P) \quad \text{is continuous and strictly increasing at } x = c_{I(P)}(1-\alpha, P). \tag{28}$$

Note that the continuity condition in (28) is satisfied if the $|I(P)|$ univariate marginal distributions of $G_{I(P)}(P)$ are continuous. Also, the strictly increasing assumption can be weakened as well, but it holds in all known examples where the continuity assumption holds, because typical limit distributions are Gaussian, chi-squared, and so on. Actually, the strictly increasing assumption can be removed entirely (see remark 1.2.1 of Politis, Romano, and Wolf 1999).

Assumption A2. Under P, $G_{n,I(P)}(P)$ converges weakly to a point mass at $\mathbf{d} = \mathbf{d}(P)$, where $\mathbf{d} = (d_1(P), \ldots, d_{|I(P)|}(P))$ is a vector of $|I(P)|$ components. [In the case where $d_j(P) = -\infty$, we mean that $T_{n,j}$ converges in probability under P to $-\infty$.]

Now we prove a basic result that can be applied to several resampling or asymptotic methods to approximate critical values. Consider the stepdown method presented in Algorithm 1, with $c_{n,K}(1-\alpha)$ replaced by some estimates $\hat{c}_{n,K}(1-\alpha)$. We consider some concrete choices later.

Theorem 4. (a) Fix P and suppose that Assumption A1 holds, so that (27) and (28) hold. Assume that the estimated critical values $\hat{c}_{n,K}(1-\alpha)$ satisfy for any $K \supset I(P)$, the estimates $\hat{c}_{n,K}(1-\alpha)$ are bounded below by $c_{I(P)}(1-\alpha)$; by this we mean, for any $\epsilon > 0$, that

$$\hat{c}_{n,K}(1-\alpha) \geq c_{I(P)}(1-\alpha) - \epsilon \quad \text{with probability } \to 1. \tag{29}$$

Then, $\limsup_n \mathrm{FWE}_P \leq \alpha$.

(b) Fix P and suppose that Assumption A1 holds. Assume that the estimated critical values are monotone in the sense that

$$\hat{c}_{n,K}(1-\alpha) \geq \hat{c}_{n,I}(1-\alpha) \quad \text{whenever } I \subset K. \tag{30}$$

Then (29) holds for all $K \supset I(P)$ if it holds in the special case where $K = I(P)$. Therefore, if Assumption A1 and the monotonicity condition (30) hold, and for any $\epsilon > 0$,

$$\hat{c}_{n,I(P)}(1-\alpha) \geq c_{I(P)}(1-\alpha) - \epsilon \quad \text{with probability } \to 1, \tag{31}$$

then $\limsup_n \mathrm{FWE}_P \leq \alpha$.

(c) Fix P and suppose that Assumption A2 holds. Also assume the monotonicity condition (30). If, for some $\epsilon > 0$,

$$\hat{c}_{n,I(P)}(1-\alpha) > \max\{d_j(P) : j \in I(P)\} + \epsilon \quad \text{with probability } \to 1, \tag{32}$$

then $\limsup_n \mathrm{FWE}_P = 0$.

Note that Assumption A1 implies that

$$c_{n,I(P)}(1-\alpha) \to c_{I(P)}(1-\alpha) \quad \text{as } n \to \infty.$$

In part (a) of Theorem 4, we replace the monotonicity requirement of Theorem 3 by a weak asymptotic monotonicity requirement (29).

In general, the point of Theorem 4 is that $\limsup_n \mathrm{FWE}_P \leq \alpha$ regardless of whether the convergence of the null hypotheses satisfies Assumption A1 or Assumption A2, at least under reasonable behavior of the estimated critical values. Moreover, we show that the monotonicity condition (30) assumed in parts (b) and (c) hold generally for some construction based on the bootstrap and subsampling. Therefore, the crux of proving strong control requires that the estimated critical values satisfy (31); that is, the critical value for testing the intersection

hypothesis $H_{I(P)}$ is consistent in that it leads to a test that asymptotically controls the probability of a type 1 error. In other words, the problem is essentially reduced to the problem of estimating the critical value for a single (intersection) test without having to worry about the multiple-testing issue of controlling the FWE. Thus the problem of controlling the FWE is reduced to the problem of controlling the type 1 error of a single test. We clarify this further for specific choices of estimates of the critical values.

Before applying Theorem 4(b) and 4(c), which assumes monotonicity of critical values, we demonstrate consistency without the assumption of monotonicity. In this regard, a simple alternative to Theorem 4(a) is the following.

Theorem 5. Fix P and suppose that Assumption A1 holds. Suppose that (29) holds for $K = I(P)$; that is, for any $\epsilon > 0$,

$$\hat{c}_{n,I(P)}(1-\alpha) \geq c_{I(P)}(1-\alpha) - \epsilon \quad \text{with probability } \to 1. \tag{33}$$

Further, suppose that the test is consistent in the sense that for any hypothesis H_j with $j \notin I(P)$, the probability of rejecting H_j by the stepdown procedure tends to 1. This happens, for example, if the critical values $\hat{c}_{n,K}$ are bounded in probability while $T_{n,j} \to \infty$ if $j \notin I(P)$. Then $\limsup_n \mathrm{FWE}_P \leq \alpha$.

Example 8 (Example 6, revisited). In the setup of Example 6, suppose that the observations are real valued, and consider a test of H_j based on

$$T_{n,j} = n^{1/2}|\bar{X}_j - \bar{X}_{k+1}|,$$

where $\bar{X}_j = n_j^{-1}\sum_i X_{i,j}$. Suppose that we use the permutation test where at stage j for testing H_{K_j}, only permutations of observations $X_{i,j}$ with $j \in K$ and $X_{i,k+1}$ are used. Assume that $n_j/n \to \lambda_j \in (0,1)$. Let $\mu(P_j)$ denote the true mean of P_j, assumed to exist; also assume the variance of P_i is finite. Then Theorem 5 applies to any P for which if $j \notin I(P)$, then $\mu(P_j) \neq \mu(P_{k+1})$ (which, of course, is not the same as $P_j \neq P_{k+1}$). Indeed, $T_{n,j} \to \infty$ in probability then. Also, using arguments similar to those of Romano (1990), $\hat{c}_{n,K}(1-\alpha)$ is bounded in probability for any K, because asymptotically it behaves like the $1-\alpha$ quantile of the maximum of $|K|$ normal variables. Therefore, asymptotic control of the FWE persists. However, if the distributions differ but the means are the same, then the test statistic should be designed to capture arbitrary differences in distribution, such as a two-sample Kolmogorov–Smirnov test statistic (unless one really wants to pick up just differences in the mean, but then the null hypothesis should reflect this.)

4.2 A Bootstrap Construction

We now specialize a bit and develop a concrete construction based on the bootstrap. For now, we suppose that hypothesis H_j is specified by $\{P : \theta_j(P) \leq 0\}$ for some real-valued parameter θ_j. Suppose that $\hat{\theta}_{n,j}$ is an estimate of θ_j. Also, let $T_{n,j} = \tau_n\hat{\theta}_{n,j}$ for some nonnegative (nonrandom) sequence $\tau_n \to \infty$. The sequence τ_n is introduced for asymptotic purposes so that a limiting distribution for $\tau_n\hat{\theta}_{n,j}$ exists when $\theta_j(P) = 0$.

Remark 4. Typically, $\tau_n = n^{1/2}$. Also, it is possible to let τ_n vary with the hypothesis j. Extensions to cases where τ_n depends on P are also possible, using ideas of Bertail, Politis, and Romano (1999).

The bootstrap method relies on its ability to approximate the joint distribution of $\{\tau_n[\hat{\theta}_{n,j} - \theta_j(P)] : j \in K\}$, whose distribution we denote by $J_{n,K}(P)$. We assume that the normalized estimates satisfy the following:

Assumption B1(a). $J_{n,I(P)}(P) \xrightarrow{L} J_{I(P)}(P)$, a nondegenerate limit law.

Let $L_{n,K}(P)$ denote the distribution under P of $\max\{\tau_n[\hat{\theta}_{n,j} - \theta_j(P)] : j \in K\}$, with corresponding distribution function $L_{n,K}(x,P)$ and α-quantile

$$b_{n,K}(\alpha,P) = \inf\{x : L_{n,K}(x,P) \geq \alpha\}.$$

Assumption B1 implies that $L_{n,I(P)}(P)$ has a limiting distribution $L_{I(P)}(P)$.

We further assume the following.

Assumption B1(b). $L_{J(P)}(P)$ is continuous and strictly increasing on its support.

Under Assumption B1, it follows that when $K = J(P)$,

$$b_{n,K}(1-\alpha,P) \to b_K(1-\alpha,P), \tag{34}$$

where $b_K(\alpha,P)$ is the α-quantile of the limiting distribution $L_K(P)$.

Assume that Assumption B1 holds. If P satisfies at least one $\theta_j(P)$ is exactly 0, then Assumption A1 holds. On the other hand, if P satisfies all $\theta_j(P) < 0$ among the $\theta_j(P)$ that are ≤ 0, then Assumption A2 holds. Indeed, if $\tau_n(\hat{\theta}_{n,j} - \theta_j(P))$ converges to a limit law and $\tau_n\theta_j(P) \to -\infty$, then $\tau_n\hat{\theta}_{n,j} \to -\infty$ in probability.

Let $\hat{Q}_n$ be some estimate of P. For iid data, $\hat{Q}_n$ is typically taken to be the empirical distribution, or possibly a smoothed version. For time series or dependent-data situations, block bootstrap methods should be used (see Lahiri 2003). Then a nominal $(1-\alpha)$-level bootstrap confidence region for the subset of parameters $\{\theta_j(P) : j \in K\}$ is given by

$$\left\{(\theta_j : j \in K) : \max_{j \in K} \tau_n[\hat{\theta}_{n,j} - \theta_j] \leq b_{n,K}(1-\alpha,\hat{Q}_n)\right\}$$
$$= \left\{(\theta_j : j \in K) : \theta_j \geq \hat{\theta}_{n,j} - \tau_n^{-1} b_{n,K}(1-\alpha,\hat{Q}_n)\right\}.$$

So a value of 0 for $\theta_j(P)$ falls outside the region iff $\tau_n\hat{\theta}_{n,j} > b_{n,K}(1-\alpha,\hat{Q}_n)$. By the usual duality of confidence sets and hypothesis tests, this suggests the use of the critical value

$$\hat{c}_{n,K}(1-\alpha) = b_{n,K}(1-\alpha,\hat{Q}_n), \tag{35}$$

at least if the bootstrap is a valid asymptotic approach for confidence region construction.

Note that, regardless of asymptotic behavior, the monotonicity assumption (30) is always satisfied for the choice (35). Indeed, for any Q and if $I \subset K$, $b_{n,I}(1-\alpha,Q)$ is the $1-\alpha$ quantile under Q of the maximum of $|I|$ variables, whereas $b_{n,K}(1-\alpha,Q)$ is the $1-\alpha$ quantile of these same $|I|$ variables together with $|K| - |I|$ variables.

Therefore, to apply Theorem 4 to conclude $\limsup_n \mathrm{FWE}_P \leq \alpha$, it is now only necessary to study the asymptotic behavior of $b_{n,K}(1-\alpha,\hat{Q}_n)$ in the case where $K = J(P)$. For this, we further assume the usual conditions for bootstrap consistency when testing the *single* hypothesis that $\theta_j(P) \leq 0$ for all $j \in J(P)$; that is, we assume that the bootstrap consistently estimates the joint distribution of $\tau_n[\hat{\theta}_{n,j} - \theta_j(P)]$ for $j \in I(P)$. Specifically, consider the following.

Assumption B2. For any metric ρ metrizing weak convergence on $\mathbb{R}^{|J(P)|}$,

$$\rho\big(J_{n,I(P)}(P), J_{n,I(P)}(\hat{Q}_n)\big) \xrightarrow{P} 0.$$

Theorem 6. Fix P satisfying Assumption B1. Let $\hat{Q}_n$ be an estimate of P satisfying Assumption B2. Consider the stepdown method in Algorithm 1 with $c_{n,K}(1-\alpha)$ replaced by $b_{n,K}(1-\alpha, \hat{Q}_n)$.

(a) Then $\limsup_n \mathrm{FWE}_P \le \alpha$.

(b) Suppose that Assumptions B_1 and B_2 hold when $I(P)$ is replaced by any subset K. If P is such that $j \notin I(P)$, [i.e., H_j is false and $\theta_j(P) > 0$], then the probability that the stepdown method rejects H_j tends to 1.

Example 9 (Continuation of Example 7). The analysis of sample correlations is a special case of the smooth function model studied by Hall (1992), and the bootstrap approach is valid for such models.

Remark 5. The main reason why the bootstrap works here can be traced to the simple result of Theorem 3. By resampling from a fixed distribution, the bootstrap approach, generates monotone critical values. Therefore, because we know how to construct valid bootstrap tests for each intersection hypothesis, this leads to valid multiple tests. But we learn more. If we use a bootstrap approach such that each intersection test has a rejection probability equal to $\alpha + O(\epsilon_n)$, then we can also deduce $\limsup_n \mathrm{FWE}_P \le \alpha + O(\epsilon_n)$. In other words, if a bootstrap method has good performance for the construction of single tests, then this translates into good performance of the bootstrap for constructing stepdown multiple tests.

Remark 6. The bootstrap can also give dramatic finite-sample gains by accommodating nonnormalities, even when the test statistics are independent (see, e.g., Westfall and Young 1993, p. 162; Westfall and Wolfinger 1997).

Remark 7. Typically, the asymptotic behavior of a test procedure when P is true will satisfy that it is consistent in the sense that all false hypotheses will be rejected with probability tending to 1 (as is the case under Thm. 6). However, we can also study the behavior of procedures against contiguous alternatives so that not all false hypotheses are rejected with probability tending to 1 under such sequences. But of course, if alternative hypotheses are in some sense close to their respective null hypotheses, then the procedures will typically reject even fewer hypotheses, and so the limiting probability of any false rejection under a sequence of contiguous alternatives should then be bounded by α.

Remark 8. In addition to constructing tests that control the FWE, we typically would like to choose test statistics that lead to procedures that are balanced in the sense that all tests have about the same power. As argued by Beran (1988a), Tu and Zhou (2000), and Rogers and Hsu (2001), balance can be desirable. Alternatively, lack of balance may be desirable so that certain tests are given more weight (see Westfall and Young 1993, p. 162; Westfall and Wolfinger 1997). Although the goal of this article has been the evaluation of significance while maintaining strong control based on given test statistics, achieving balance is best handled by an appropriate choice of test statistics. For example, transforming test statistics to p-values and then using the negative p-values as the basic statistics will lead to better balance. Quite generally, Beran's prepivoting transformation can lead to balance (see Beran 1988a,b). The assumptions of our theorem must then hold for the transformed test statistics. Alternatively, balance sometimes can be achieved by studentization. The construction developed in this section can be extended to the case of studentized test statistics. The details are straightforward and left to the reader.

We now briefly consider the two-sided case. Suppose that H_j specifies $\theta_j(P) = 0$ against the alternative $\theta_j(P) \neq 0$. Let $L'_{n,K}(P)$ denote the distribution under P of $\max\{\tau_n|\hat{\theta}_{n,j} - \theta_j(P)| : j \in K\}$ with corresponding distribution function $L'_{n,K}(x, P)$ and α-quantile

$$b'_{n,K}(\alpha, P) = \inf\{x : L'_{n,K}(x, P) \ge \alpha\}.$$

Accordingly, $L'_K(P)$ denotes the limiting distribution of $L'_{n,K}(P)$. Finally, let $T'_{n,j} = \tau_n|\hat{\theta}_{n,j}|$.

Theorem 7. Fix P satisfying Assumption B1, but with $L_{I(P)}(P)$ in Assumption B1(b) replaced by $L'_{I(P)}(P)$. Let $\hat{Q}_n$ be an estimate of P satisfying Assumption B2. Consider the stepdown method in Algorithm 1 using the test statistics $T'_{n,j}$ and with $c_{n,K}(1-\alpha)$ replaced by $b'_{n,K}(1-\alpha, \hat{Q}_n)$.

(a) Then $\limsup_n \mathrm{FWE}_P \le \alpha$.

(b) Suppose that Assumptions B1 and B2 hold when $I(P)$ is replaced by any subset K (and L is replaced by L'). If P is such that $j \notin I(P)$ [i.e., H_j is false and $\theta_j(P) \neq 0$] then the probability that the stepdown method rejects H_j tends to 1.

(c) Moreover, if the foregoing algorithm rejects H_j and it is declared that $\theta_j > 0$ when $\hat{\theta}_{n,j} > 0$, then the probability of making a type 3 error [i.e., of declaring that $\theta_j(P)$ is positive when it is negative or declaring it negative when it is positive] tends to 0.

An alternative approach to the two-sided case is to balance the tails of the bootstrap distribution of the original estimates (without the absolute values) separately. An analogous result would hold. The comparison of these approaches in the case of a single test was made by Hall (1992).

Theorem 7(c) shows that the directional error is asymptotically negligible. It would be more interesting to obtain finite-sample results and to study the behavior of the directional error under contiguous alternatives so that the problem is no longer asymptotically degenerate; future work will consider these problems. (For references to the literature on controlling the directional error as well as some finite-sample results, see Finner 1999.)

4.3 A Subsampling Construction

In this section we present an alternative construction that applies under weaker conditions than the bootstrap. We now assume that we have available an iid sample $\mathbf{X}_1, \ldots, \mathbf{X}_n$ from P, and that $T_{n,j} = T_{n,j}(\mathbf{X}_1, \ldots, \mathbf{X}_n)$ is the test statistic that we wish to use for testing H_j. To describe the test construction, fix a positive integer $b \le n$ and let $Y_1, \ldots, Y_{N_n}$ be equal to the $N_n = \binom{n}{b}$ subsets of $\{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$, ordered in any fashion. Let $T^{(i)}_{b,j}$ be equal to the statistic $T_{b,j}$ evaluated at the dataset Y_i. Then, for

any subset $K \subset \{1, \ldots, k\}$, the joint distribution of $(T_{n,j} : j \in K)$ can be approximated by the empirical distribution of the $\binom{n}{b}$ values $(T_{b,j}^{(i)} : j \in K)$. In other words, for $\mathbf{x} \in \mathbb{R}^k$, the true joint cdf of the test statistics evaluated at $\mathbf{x}$,

$$G_{n,\{1,\ldots,k\}}(\mathbf{x}, P) = P\{T_{n,1} \le x_1, \ldots, T_{n,k} \le x_k\},$$

is estimated by the subsampling distribution

$$\hat{G}_{n,\{1,\ldots,k\}}(\mathbf{x}) = \binom{n}{b}^{-1} \sum_i I\{T_{b,1}^{(i)} \le x_1, \ldots, T_{b,k}^{(i)} \le x_k\}. \quad (36)$$

Note that the marginal distribution of any subset $K \subset \{1, \ldots, k\}$, $G_{n,K}(P)$, is then approximated by the marginal distribution induced by (36) on that subset of variables. So, $\hat{G}_{n,K}$ refers to the empirical distribution of the values $(T_{n,j}^{(i)} : j \in K)$. (In essence, one need only estimate one joint sampling distribution for all the test statistics, because this then induces that of any subset, even though we are not assuming anything like subset pivotality.)

Similarly, the estimate of the whole joint distribution of test statistics induces an estimate for the distribution of the maximum of test statistics. Specifically, $H_{n,K}(P)$ is estimated by the empirical distribution $\hat{H}_{n,K}(x)$ of the values $\max(T_{n,j}^{(i)} : j \in K)$, that is,

$$\hat{H}_{n,K}(x) = \binom{n}{b}^{-1} \sum_i I\{\max(T_{b,j}^{(i)} : j \in K) \le x\}.$$

Also, let

$$\hat{c}_{n,K}(1-\alpha) = \inf\{x : \hat{H}_{n,K}(x) \ge 1-\alpha\}$$

denote the estimated $1-\alpha$ quantile of the maximum of test statistics $T_{n,j}$ with $j \in K$.

Note the monotonicity of the critical values; for $I \subset K$,

$$\hat{c}_{n,K}(1-\alpha) \ge \hat{c}_{n,I}(1-\alpha), \quad (37)$$

and so the monotonicity assumption in Theorem 4 holds [also compare with (4)].

This leads us to consider the idealized stepdown algorithm with $c_{n,K}(1-\alpha, P)$ replaced by the estimates $\hat{c}_{n,K}(1-\alpha)$. The following result proves consistency and strong control of this subsampling approach. Note in particular that Assumption B2 is not needed here at all, a reflection of the fact that the bootstrap requires much stronger conditions for consistency (see Politis et al. 1999). Also notice that we do not even need to assume that there exists a P for which all hypotheses are true.

Theorem 8. Suppose that Assumption A1 holds. Let $b/n \to 0$, $\tau_b/\tau_n \to 0$, and $b \to \infty$.

(a) The subsampling approximation satisfies

$$\rho\big(\hat{G}_{n,I(P)}, G_{n,I(P)}(P)\big) \xrightarrow{P} 0, \quad (38)$$

for any metric ρ metrizing weak convergence on $\mathbb{R}^{|I(P)|}$.

(b) The subsampling critical values satisfy

$$\hat{c}_{n,I(P)}(1-\alpha) \xrightarrow{P} c_{I(P)}(1-\alpha). \quad (39)$$

(c) Therefore, using Algorithm 1 with $c_{n,K}(1-\alpha, P)$ replaced by the estimates $\hat{c}_{n,K}(1-\alpha)$ results in $\limsup_n \mathrm{FWE}_P \le \alpha$.

Example 10 (Cube root asymptotics). Kim and Pollard (1990) showed that a general class of M-estimators converge at rate $\tau_n = n^{1/3}$ to a nonnormal limiting distribution. As a result, inconsistency of the bootstrap typically follows. Rodríguez-Poo, Delgado, and Wolf (2001) demonstrated the consistency of the subsampling method for constructing hypothesis tests for a single null hypothesis. By similar arguments, the validity of the subsampling construction of Theorem 8 in the context of cube root asymptotics can be established.

The foregoing approach can be extended to dependent data. For example, if the data form a stationary sequence, then we would consider only the $n-b+1$ subsamples of the form $(\mathbf{X}_i, \mathbf{X}_{i+1}, \ldots, \mathbf{X}_{i+b-1})$. Generalizations for nonstationary time series, random fields, and point processes were further treated by Politis et al. (1999).

5. TWO SIMULATION STUDIES

5.1 Testing Means

This section presents a small simulation study in the context of testing population means. We generate random vectors $\mathbf{X}_1, \ldots, \mathbf{X}_{100}$ from a k-dimensional multivariate normal distribution with mean vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)'$. The values of k are $k = 10$ and $k = 40$. Each null hypothesis is $H_j : \theta_j \le 0$, and each alternative hypothesis is one-sided. We apply the stepdown bootstrap construction of Section 4.2, resampling from the empirical distribution. In the spirit of Remark 8, we use the studentized test statistics $T_{n,j} = \sqrt{n}\bar{X}_j/s_j$, where $\bar{X}_j$ and s_j are the usual sample average and sample standard deviation of jth sample. In addition, we also include the Holm method in the study. The nominal FWE levels are $\alpha = .05$ and $\alpha = .1$. Performance criteria are the empirical FWE and the (average) number of false hypotheses that are rejected.

We consider three scenarios for the mean vector of the multivariate normal distribution. In the first scenario, all means θ_j are equal to 0. In the second scenario, half of the means are equal to 0 and the other half are equal to .25. In the third scenario, all means are equal to .25.

We consider three scenarios for the covariance matrix of the multivariate normal distribution. In the first scenario, the covariance matrix is the identity matrix. In the second scenario, all of the variances are equal to 1 and all of the correlations are equal to $\rho = .5$. In the third scenario, all of the variances are equal to 1 and all of the correlations are equal to $\rho = .9$. We would expect our stepwise method to perform similarly to that of Holm in the first scenario but to reject more false null hypotheses in the latter two scenarios.

Tables 1 and 2 report the results based on 10,000 repetitions. The number of bootstrap resamples is $B = 1{,}000$. The results demonstrate the good control of the FWE in finite samples and increased power of the stepdown method compared with the Holm method in cases of a positive common correlation ρ.

Note that the FWE control of the Holm method for the case where $\rho = 0$ deteriorates somewhat when the number of hypotheses tested increases from $k = 10$ to $k = 40$, but that this does not happen with the stepdown method. The reason for this behavior of the Holm method is that individual p-values are computed using the asymptotic standard normal distribution of

Table 1. Empirical FWEs and Average Number of False Hypotheses Rejected for Both the Holm Method and the General Stepdown Construction of Section 4.2

Level α	FWE (Holm)	FWE (stepdown)	Rejected (Holm)	Rejected (stepdown)
All $\theta_j = 0$, all $\rho = 0$				
5	5.8	5.1	.0	.0
10	10.5	9.8	.0	.0
All $\theta_j = 0$, all $\rho = .5$				
5	3.9	5.2	.0	.0
10	7.2	10.3	.0	.0
All $\theta_j = 0$, all $\rho = .9$				
5	1.9	5.2	.0	.0
10	3.3	10.1	.0	.0
Half of the $\theta_j = .25$, all $\rho = 0$				
5	4.1	3.6	2.5	2.5
10	8.2	7.9	3.0	3.1
Half of the $\theta_j = .25$, all $\rho = .5$				
5	3.8	5.1	2.5	2.7
10	7.4	10.0	3.1	3.3
Half of the $\theta_j = .25$, all $\rho = .9$				
5	2.5	5.0	2.5	3.4
10	4.5	10.0	3.0	3.9
All $\theta_j = .25$, all $\rho = 0$				
5	.0	.0	5.8	5.8
10	.0	.0	7.3	7.3
All $\theta_j = .25$, all $\rho = .5$				
5	.0	.0	5.8	6.1
10	.0	.0	7.0	7.5
All $\theta_j = .25$, all $\rho = .9$				
5	.0	.0	5.7	7.1
10	.0	.0	6.8	8.2

NOTE: The nominal levels are $\alpha = 5\%$ and $\alpha = 10\%$. Observations are iid multivariate normal, the dimension is $k = 10$, and the number of observations is $n = 100$. The number of repetitions is 10,000 per scenario, and the number of bootstrap resamples is $B = 1{,}000$.

Table 2. Empirical FWEs and Average Number of False Hypotheses Rejected for Both the Holm Method and the General Stepdown Construction of Section 4.2

Level α	FWE (Holm)	FWE (stepdown)	Rejected (Holm)	Rejected (stepdown)
All $\theta_j = 0$, all $\rho = 0$				
5	6.1	4.8	.0	.0
10	11.6	10.0	.0	.0
All $\theta_j = 0$, all $\rho = .5$				
5	3.8	4.9	.0	.0
10	6.4	10.0	.0	.0
All $\theta_j = 0$, all $\rho = .9$				
5	1.0	5.0	.0	.0
10	1.6	9.9	.0	.0
Half of the $\theta_j = .25$, all $\rho = 0$				
5	3.4	3.2	6.4	6.3
10	6.9	6.7	8.2	8.1
Half of the $\theta_j = .25$, all $\rho = .5$				
5	3.5	4.6	6.6	7.6
10	6.3	9.4	8.3	10.0
Half of the $\theta_j = .25$, all $\rho = .9$				
5	1.1	5.3	6.7	12.0
10	2.2	10.1	8.4	14.6
All $\theta_j = .25$, all $\rho = 0$				
5	.0	.0	14.1	13.5
10	.0	.0	18.4	18.1
All $\theta_j = .25$, all $\rho = .5$				
5	.0	.0	15.3	17.5
10	.0	.0	19.6	23.2
All $\theta_j = .25$, all $\rho = .9$				
5	.0	.0	15.6	25.1
10	.0	.0	19.2	30.3

NOTE: The nominal levels are $\alpha = 5\%$ and $\alpha = 10\%$. Observations are iid multivariate normal, the dimension is $k = 40$, and the number of observations is $n = 100$. The number of repetitions is 10,000 per scenario and the number of bootstrap resamples is $B = 1{,}000$.

the t-statistic under the null. Because the true distribution under the null is t_{99}, the p-values are somewhat anticonservative in finite samples, and when k increases from 10 to 40, this effect is apparently magnified. (Of course, using the t_{99} distribution instead to compute individual p-values would correspond to knowing the parametric nature of the underlying probability mechanism, which is not realistic.)

Further note that relative advantage in terms of power of the stepdown method does not seem to diminish when the number of hypotheses tested increases from $k = 10$ to $k = 40$. For example, consider the case where $\alpha = .1$ and $\rho = .9$. When $k = 10$, the stepdown method on average rejects about 20% more false hypotheses compared with the Holm method. When $k = 40$, the improvement is about 50%. Of course, this is an observation restricted to the particular data-generating mechanism used in our simulation study and should not be interpreted as a general theoretical statement.

5.2 Testing Correlations

This section presents a small simulation study in the context of Example 7. We generate random vectors $\mathbf{X}_1, \ldots, \mathbf{X}_n$ from a 10-dimensional multivariate normal distribution. Hence there are a total of $k = \binom{10}{2} = 45$ pairwise correlations to test. The values for the sample size are $n = 50$ and $n = 100$. Each null hypothesis is $H_{i,j} : \rho_{i,j} = 0$ and each alternative hypothesis is two-sided. We apply the stepdown bootstrap construction of Section 4.2, resampling from the empirical distribution. The test statistics are given by $T_{n,i,j} = \sqrt{n}\hat{\rho}_{i,j}$, where $\hat{\rho}_{i,j}$ is the usual sample correlation between the ith sample and the jth sample. As a special case, we also look at the corresponding single-step method. The nominal FWE levels are $\alpha = .05$ and $\alpha = .1$; performance criteria are the empirical FWE and the (average) number of false hypotheses that are rejected.

We consider three scenarios. In the first scenario, all correlations are equal to 0. In the second scenario, all $\rho_{1,j}$ are equal to .3, for $j = 2, \ldots, 10$, and the remaining correlations are equal to 0. In the third scenario, all correlations are equal to .3.

Table 3 reports the results based on 10,000 repetitions. The number of bootstrap resamples is $B = 1{,}000$. The results demonstrate the good control of the FWE in finite samples and the increased power of the stepdown method compared with the single-step method.

6. EMPIRICAL APPLICATION

Westfall and Young (1993, example 6.4) applied a multiple-testing method for 10 pairwise correlations. Each null hypothesis is that corresponding pairwise population correlation is equal to 0; and each alternative hypothesis is two-sided. (See their example 6.4 for the details of the rea dataset.) Westfall and Young (1993) carried out a bootstrap multiple test under the assumption of complete independence. As they admitted, this is a conservative approach in general. Instead, we apply the stepdown bootstrap construction of Section 4.2, resampling from the empirical distribution. For each null hypothesis, the

Table 3. Empirical FWEs and Average Number of False Hypotheses Rejected for Both the Single-Step Construction and the General Stepdown Construction of Section 4.2

Level α	FWE (single-step)	FWE (stepdown)	Rejected (single-step)	Rejected (stepdown)
All $\rho_{i,j} = 0$, $n = 50$				
5	3.5	3.5	.0	.0
10	8.2	8.2	.0	.0
All $\rho_{1,j} = .3$ and remaining $\rho_{i,j} = 0$, $n = 50$				
5	3.0	3.0	.96	.97
10	7.6	7.6	1.5	1.6
All $\rho_{i,j} = .3$, $n = 50$				
5	.0	.0	6.6	7.1
10	.0	.0	10.2	11.2
All $\rho_{i,j} = 0$, $n = 100$				
5	4.4	4.4	.0	.0
10	9.7	9.7	.0	.0
All $\rho_{1,j} = .3$ and remaining $\rho_{i,j} = 0$, $n = 100$				
5	3.9	4.0	3.7	3.8
10	8.7	9.0	4.5	4.6
All $\rho_{i,j} = .3$, $n = 100$				
5	.0	.0	21.9	25.4
10	.0	.0	26.5	30.9

NOTE: The nominal levels are $\alpha = 5\%$ and $\alpha = 10\%$. Observations are iid multivariate normal, the number of observations is $n = 50$ and $n = 100$, and the number of pairwise correlations is $k = 45$. The number of repetitions is 10,000 per scenario, and the number of bootstrap resamples is $B = 1{,}000$.

stepdown construction yields an adjusted p-value; it is given by the smallest FWE level α at which the construction rejects this particular hypothesis.

Table 4 compares the adjusted p-values of Westfall and Young (1993) to ours. The conservativeness of Westfall and Young's method can be clearly appreciated.

7. CONCLUDING REMARKS

We have shown how computationally feasible stepdown methods can be constructed to control the FWE in a fair amount of generality. Further study is needed to study the control of directional errors, and future work will focus on a similar treatment for stepup procedures. We also would like to extend our results to show how resampling can be used to estimate the dependence structure of the test statistics to obtain improved methods that control the false discovery rate of Benjamini and Hochberg (1995). Some results were obtained by Benjamini and Yekutieli (2001), but these authors also assumed the subset pivotality condition. By extending our work, we hope to avoid such conditions.

Table 4. Sample Correlations and p-Values for the Data of Example 6.4 of Westfall and Young (1993)

Variables	Sample correlation	Raw p-value	W–Y p-value	Step p-value
(SATdev, % Black)	−.5089	.0002	.0019	.0016
(Salary, Crime)	.4902	.0003	.0030	.0028
(% Black, Crime)	.4844	.0004	.0036	.0034
(SATdev, S/T Ratio)	−.3864	.0061	.0404	.0346
(SATdev, Crime)	−.3033	.0341	.1843	.1483
(S/T Ratio, Crime)	.2290	.1135	.4485	.3921
(S/T Ratio, % Black)	.1732	.2341	.6474	.5986
(SATdev, Salary)	.0980	.5030	.8753	.8572
(Salary, % Black)	−.0354	.8090	.9641	.9645
(S/T Ratio, Salary)	.0045	.9754	.9759	.9761

NOTE: "W–Y p-value" denotes the adjusted p-value of Westfall and Young; "Step p-value" denotes the adjusted bootstrap p-value obtained from the stepdown construction of Section 4.2 (based on $B = 5{,}000$ bootstrap resamples).

APPENDIX: PROOFS

Proof of Theorem 1

Consider the event where a true hypothesis is rejected, so that for some $j \in I(P)$, hypothesis H_j is rejected. Let $\hat{j}$ be the (random) smallest index j in the algorithm where this occurs, so that

$$T_{n,r_{\hat{j}}} > c_{n,K_{\hat{j}}}(1-\alpha). \tag{A.1}$$

Because $K_{\hat{j}} \supset I(P)$, assumption (5) implies that

$$c_{n,K_{\hat{j}}}(1-\alpha) \geq c_{n,I(P)}(1-\alpha) \geq c_{n,I(P)}(1-\alpha, P),$$

and so

$$T_{n,r_{\hat{j}}} > c_{n,I(P)}(1-\alpha, P).$$

Furthermore, by definition of $\hat{j}$,

$$T_{n,r_{\hat{j}}} = \max(T_{n,j},\ j \in K_{\hat{j}}) = \max\bigl(T_{n,j},\ j \in I(P)\bigr),$$

and so the event that a false rejection occurs under P implies that

$$\max\bigl(T_{n,j}, j \in I(P)\bigr) > c_{n,I(P)}(1-\alpha, P). \tag{A.2}$$

Therefore, the probability of a type 1 error is bounded above by the probability of the event (A.2), which by definition has probability bounded above by α. The proof of part (b) is obvious, because the procedure becomes more conservative. The proof of (c) holds by the proof of (a) on replacing the constants $c_{n,K_{\hat{j}}}(1-\alpha)$ by $d_{n,K_{\hat{j}}}(1-\alpha)$.

Proof of Corollary 1

We verify the conditions for $d_{n,K_j}(1-\alpha)$ when $d_{n,K_j}(1-\alpha) = c^*_{n,K_j}(1-\alpha)$ in Theorem 1(b) and 1(c). Clearly,

$$c^*_{n,K}(1-\alpha) \geq c_{n,I}(1-\alpha).$$

Also, for $K \supset I(P)$,

$$\begin{aligned} c^*_{n,K}(1-\alpha) &= \max\{c_{n,J}(1-\alpha) : J \subset K\} \\ &\geq \max\{c_{n,J}(1-\alpha) : J \subset I(P)\} \\ &= c^*_{n,I(P)}(1-\alpha), \end{aligned}$$

and so (7) holds.

Proof of Theorem 2

To prove (a), let $\hat{j}$ be the smallest (random) index j such that $T_{n,r_{\hat{j}}} > \tilde{c}_{n,K_{\hat{j}}}(1-\alpha)$. But $K_{\hat{j}} \supset I(P)$, and so

$$\tilde{c}_{n,K_{\hat{j}}}(1-\alpha) \geq \tilde{c}_{n,I(P)}(1-\alpha) \geq c_{n,I(P)}(1-\alpha, P).$$

Thus the event that a false rejection occurs under P implies that

$$\max\bigl(T_{n,j},\ j \in I(P)\bigr) > c_{n,I(P)}(1-\alpha, P), \tag{A.3}$$

which has probability bounded by α. The proof of (b) is obvious, because the procedure becomes more conservative.

Proof of Theorem 3

As in the argument of Theorem 1, the event a false rejection occurs is the event

$$\max\{T_{n,j} : j \in I(P)\} > \hat{c}_{n,K_{\hat{j}}}(1-\alpha), \tag{A.4}$$

where $\hat{j}$ is the smallest (random) index where a false rejection occurs. Because $K_{\hat{j}} \supset I(P)$,

$$\hat{c}_{n,K_{\hat{j}}}(1-\alpha) \geq \hat{c}_{n,I(P)}(1-\alpha), \tag{A.5}$$

and so (a) follows. Part (b) follows immediately from (a).

Proof of Theorem 4

As in the proofs of Theorems 1 and 3, namely (A.4), it suffices to show that

$$\limsup_n P\{\max\{T_{n,j}: j \in I(P)\} > \hat{c}_{n,K_j}(1-\alpha)\} \le \alpha.$$

But assumption (29) implies that

$$\hat{c}_{n,K_j}(1-\alpha) \ge c_{I(P)}(1-\alpha) - \epsilon \quad \text{with probability } \to 1.$$

Therefore, using Assumption A1, the limit superior of the probability of a false rejection is bounded above by

$$\limsup_n \mathrm{FWE}_P \le P\{\max(T_j,\ j \in I(P)) > c_{I(P)}(1-\alpha) - \epsilon\},$$

where $(T_j,\ j \in I(P))$ denote variables whose joint distribution is $G_{I(P)}(P)$. But letting $\epsilon \to 0$, the right side of the last expression becomes

$$1 - H_{I(P)}(c_{I(P)}(1-\alpha), P) = 1 - (1-\alpha) = \alpha.$$

To prove (b), because (29) holds when $K = I(P)$, then it must hold for any K containing $I(P)$, by assumption (30).

To prove (c), the probability of false rejection [i.e., the event (A.4)], is again bounded by the probability of the event

$$\max\{T_{n,j}: j \in I(P)\} > \hat{c}_{n,I(P)}(1-\alpha),$$

which converges to 0 by Assumption A2 and (32).

Proof of Theorem 5

Following the proof of Theorem 4(a), the random index $\hat{j}$ is equal to $k - |I(P)| + 1$ with probability tending to 1, and this index is no longer random. That is, with probability tending to 1, we first reject all false hypotheses and then commit a false rejection when we get to the stage at which we are testing the $|I(P)|$ true hypotheses. But then Assumption A1 and (33) allow us to conclude control of the FWE.

Proof of Theorem 6

To prove (a), fix P and assume that $\theta_j(P) = 0$ for at least one $j \in I(P)$. Then, by the comments leading up to the statement of the theorem, the conditions of Theorem 4(b) are satisfied if we can verify that

$$b_{n,I(P)}(1-\alpha, \hat{Q}_n) \xrightarrow{P} c_{I(P)}(1-\alpha).$$

But by the continuous mapping theorem, Assumption B2 implies that

$$\rho_1(L_{n,I(P)}(P), L_{n,I(P)}(\hat{Q}_n)) \xrightarrow{P} 0,$$

where ρ_1 is any metric metrizing weak convergence on $\mathbb{R}$. Furthermore, $L_{n,I(P)}(P)$ converges weakly to a continuous limit law by Assumption B1, and so

$$b_{n,I(P)}(1-\alpha, \hat{Q}_n) \xrightarrow{P} b_{I(P)}(1-\alpha, P)$$

and

$$b_{n,I(P)}(1-\alpha, P) \to b_{I(P)}(1-\alpha, P).$$

Thus it suffices to show that

$$\liminf b_{n,I(P)}(1-\alpha, P) \to c_{I(P)}(1-\alpha, P). \tag{A.6}$$

But for $\theta_j(P) \le 0$,

$$\tau_n[\hat{\theta}_{n,j} - \theta_j(P)] \ge \tau_n \hat{\theta}_{n,j} = T_n,$$

which implies that

$$b_{n,I(P)}(1-\alpha, P) \ge c_{n,I(P)}(1-\alpha, P).$$

But the right term converges to $c_{I(P)}(1-\alpha, P)$, and so (A.6) follows.

Next, assume that P has $\theta_j(P) < 0$ for all $j \in I(P)$. Then we just need to verify the conditions of Theorem 4(c). All that remains to verify is, for some $\epsilon > 0$,

$$b_{n,I(P)}(1-\alpha, \hat{Q}_n) > \max\{d_j(P): j \in I(P)\} + \epsilon,$$

with probability tending to 1. But the right side here is $-\infty$ (for any finite ϵ), so it just suffices to verify that the left side is $O_P(1)$. But by Assumption B2, it suffices to show that $b_{n,I(P)}(1-\alpha, P)$ is bounded away from $-\infty$, which follows by (34).

To prove (b), the assumptions imply that, for any $K \supset I(P)$,

$$b_{n,K}(1-\alpha, \hat{Q}_n) \xrightarrow{P} b_K(1-\alpha, P) < \infty.$$

But

$$\max(T_{n,j}: j \in K) \ge T_{n,j} = \tau_n \hat{\theta}_{n,j} \xrightarrow{P} \infty,$$

because $\hat{\theta}_{n,j} \xrightarrow{P} \theta_j(P) > 0$ and $\tau_n \to \infty$. Therefore, with probability tending to 1, for any $K \supset I(P)$,

$$\max(T_{n,j}: j \in K) \ge b_{n,K}(1-\alpha, \hat{Q}_n),$$

and the result follows.

Proof of Theorem 7

The proof is completely analogous to the proof of Theorem 6. The only additional fact needed to prove (c) is that when $\theta_j(P) > 0$, $\tau_n \hat{\theta}_{n,j} > 0$ with probability tending to 1, and similarly for $\theta_j(P) < 0$. Indeed, Assumption B1(a) implies that $\tau_n(\hat{\theta}_{n,j} - \theta_j(P))$ has a limiting distribution, which implies that $\tau_n \hat{\theta}_{n,j} \xrightarrow{P} \infty$ if $\theta_j(P) > 0$.

Proof of Theorem 8

The proof of (a) is the essential subsampling argument, which derives from (36) being a U-statistic (see Politis et al. 1999, thm. 2.6.1) where one statistic is treated, but the argument is extendable to the simultaneous estimation of the joint distribution. The result (b) follows as well. To verify (c), apply Theorem 4(b). The monotonicity requirement follows by (37), and (31) follows by (b).

[Received December 2003. Revised March 2004.]

REFERENCES

Aitken, M. (1969), "Some Tests for Correlation Matrices," *Biometrika*, 56, 443–446.

——— (1971), Correction to "Some Tests for Correlation Matrices," *Biometrika*, 58, 245.

Benjamini, Y., and Hochberg, Y. (1995), "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," *Journal of the Royal Statistical Society*, Ser. B, 57, 289–300.

Benjamini, Y., and Yekutieli, D. (2001), "The Control of the False Discovery Rate in Multiple Testing Under Dependency," *The Annals of Statistics*, 29, 1165–1188.

Beran, R. (1986), "Simulated Power Functions," *The Annals of Statistics*, 14, 151–173.

——— (1988a), "Balanced Simultaneous Confidence Sets," *Journal of the American Statistical Association*, 83, 679–686.

——— (1988b), "Prepivoting Test Statistics: A Bootstrap View of Asymptotic Refinements," *Journal of the American Statistical Association*, 83, 687–697.

Bertail, P., Politis, D., and Romano, J. (1999), "On Subsampling Estimators With Unknown Rate of Convergence," *Journal of the American Statistical Association*, 94, 569–579.

Dudoit, S., Shaffer, J., and Boldrick, J. (2003), "Multiple Hypothesis Testing in Microarray Experiments," *Statistical Science*, 18, 71–103.

Finner, H. (1999), "Stepwise Multiple Test Procedures and Control of Directional Errors," *The Annals of Statistics*, 27, 274–289.

Finner, H., and Roters, M. (1998), "Asymptotic Comparison of Step-Down and Step-Up Multiple Test Procedures Based on Exchangeable Test Statistics," *The Annals of Statistics*, 26, 505–524.

Hall, P. (1992), *The Bootstrap and Edgeworth Expansion*, New York: Springer-Verlag.

Hall, P., and Wilson, S. (1991), "Two Guidelines for Bootstrap Hypothesis Testing," *Biometrics*, 47, 757–762.
Hochberg, Y., and Tamhane, A. (1987), *Multple Comparison Procedures*, New York: Wiley.
Hoeffding, W. (1952), "The Large-Sample Power of Tests Based on Permutations of Observations," *The Annals of Mathematical Statistics*, 23, 169–192.
Holm, S. (1979), "A Simple Sequentially Rejective Multiple Test Procedure," *Scandinavian Journal of Statistics*, 6, 65–70.
Hommel, G. (1986), "Multiple Test Procedures for Arbitrary Dependence Structures," *Metrika*, 33, 321–336.
Kim, J., and Pollard, D. B. (1990), "Cube Root Asymptotics," *The Annals of Statistics*, 18, 191–219.
Lahiri, S. N. (2003), *Resampling Methods for Dependent Data*, New York: Springer-Verlag.
Lehmann, E., Romano, J. P., and Shaffer, J. (2003), "On Optimality of Stepdown and Stepup Multiple Test Procedures," Technical Report 2003-12, Stanford University, Dept. of Statistics.
Marcus, R., Peritz, E., and Gabriel, K. (1976), "On Closed Testing Procedures With Special Reference to Ordered Analysis of Variance," *Biometrika*, 63, 655–660.
Petrondas, D., and Gabriel, K. (1983), "Multiple Comparisons by Rerandomization Tests," *Journal of the American Statistical Association*, 78, 949–957.
Politis, D. N., Romano, J. P., and Wolf, M. (1999), *Subsampling*, New York: Springer-Verlag.
Rodríguez-Poo, J., Delgado, M., and Wolf, M. (2001), "Subsampling Inference in Cube Root Asymptotics With an Application to Manski's Maximum Score Estimator," *Economics Letters*, 73, 241–250.
Rogers, J., and Hsu, J. (2001), "Multiple Comparisons of Biodiversity," *Biometrical Journal*, 43, 617–625.
Romano, J. (1988), "A Bootstrap Revival of Some Nonparametric Distance Tests," *Journal of the American Statistical Association*, 83, 698–708.
——— (1989), "Bootstrap and Randomization Tests of Some Nonparametric Hypotheses," *The Annals of Statistics*, 17, 141–159.
——— (1990), "On the Behavior of Randomization Tests Without a Group Invariance Assumption," *Journal of the American Statistical Association*, 85, 686–692.
Troendle, J. (1995), "A Stepwise Resampling Method of Multiple Hypothesis Testing," *Journal of the American Statistical Association*, 90, 370–378.
Tu, W., and Zhou, X. (2000), "Pairwise Comparison of the Means of Skewed Data," *Journal of Statistical Planning and Inference*, 88, 59–74.
Westfall, P. H., and Wolfinger, R. D. (1997), "Multiple Tests With Discrete Distributions," *The American Statistician*, 51, 3–8.
Westfall, P. H., and Young, S. S. (1993), *Resampling-Based Multiple Testing: Examples and Methods for P-Value Adjustment*, New York: Wiley.
Westfall, P. H., Zaykin, D. V., and Young, S. S. (2001), "Multiple Tests for Genetic Effects in Association Studies," in *Methods in Molecular Biology: Biostatistical Methods*, Vol. 184, ed. S. Looney, Toloway, NJ: Humana Press, pp. 143–168.