Exact and Approximate Stepdown Methods For Multiple Hypothesis Testing

Joseph P. Romano*
Department of Statistics
Stanford University

Michael Wolf†
Department of Economics and Business
Universitat Pompeu Fabra

December 2003

Abstract

Consider the problem of testing k hypotheses simultaneously. In this paper, we discuss finite and large sample theory of stepdown methods that provide control of the familywise error rate (FWE). In order to improve upon the Bonferroni method or Holm's (1979) stepdown method, Westfall and Young (1993) make 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 called subset pivotality. The goal of this paper is to construct general stepdown methods that do not require such an assumption. In order to accomplish this, we take a close look at what makes stepdown procedures work, and a key component is a monotonicity requirement of critical values. By imposing such monotonicity on estimated critical values (which is not an assumption on the model but an assumption on the method), it is demonstrated that the problem of constructing a valid multiple test procedure which controls the FWE can be reduced to the problem of contructing a single test which controls the usual probability of a Type 1 error. This reduction allows us to draw upon an enormous resampling literature as a general means of test contruction.

KEY WORDS: Bootstrap, Familywise Error Rate, Multiple Testing, Permutation Test, Randomization Test, Stepdown Procedure, Subsampling.

JEL CLASSIFICATION NOS: C12, C14.

*Research supported by National Science Foundation grant DMS 010392.

†Research supported by the Spanish Ministry of Science and Technology and FEDER, grant BMF2003-03324, and by the Barcelona Economics Program of CREA.

1 Introduction

The main point of this paper is to show how computer-intensive methods can be used to construct asymptotically valid tests of multiple hypotheses under very weak conditions. The treatise by Westfall and Young (1993) takes good advantage of resampling to estimate the dependence structure of multiple test statistics in order to construct more efficient multiple testing methods. However, their methods rely heavily on the assumption of subset pivotality. Thus, the main goal of this paper is to show how to construct valid stepdown methods that do not require this assumption, while still being computationally feasible.

In Section 2, we discuss stepdown methods that control the familywise error rate in finite samples. Such methods proceed stagewise by testing an intersection hypothesis without regard to hypotheses previously rejected. However, 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.

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 constructing a critical value for a single test. 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, which is done in Section 4.

In Sections 5 and 6, we present a small simulation study and an empirical application, respectively. All proofs are collected in an appendix.

Thus, this work is a sustained essay designed to reduce the construction of stepdown methods that control the familywise error rate for multiple testing to the problem of construction of single tests that control the probability of a type 1 error, which then allows us to draw upon an enormous resampling literature.

Further 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 in order to obtain improved methods that control the false discovery rate of Benjamini and Hochberg (1995). Some results are obtained in Benjamini and Yekutieli (2001), but they also assume the subset pivotality condition. By extending our work, we hope to avoid such conditions.

2 Nonasymptotic Results

Suppose data X is generated from some unknown probability distribution P. In anticipation of asymptotic results, we may write $X = X^{(n)}$, where n typically refers to the sample size. A model assumes that P belongs to a certain family of probability distributions Ω, though 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 $P \in \bigcap_{j \in K} \omega_j$.

In this section, we tacitly assume that H_K is not empty for any subset K of $\{1, \ldots, k\}$; this is the *free combinations* condition of Holm (1979); that is, for any K, the intersection hypothesis H_K is not empty.

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 exist to approximate a critical value, such as those based on classical likelihood theory, bootstrap tests, Edgeworth expansions, permutation tests, etc. The main problem addressed in the present work is to construct a procedure that controls the familywise error rate (FWE). Recall that the familywise error rate is the probability of rejecting at least one true null hypothesis. More specifically, if P is the true probability mechanism, let $I = I(P) \subset \{1, \ldots, k\}$ denote the indices of the set of true hypotheses; that is, $i \in I$ if and only $P \in \omega_i$. The FWE is the probability under P that any H_i with $i \in I$ is rejected. To show its dependence on P, we may write $\text{FWE} = \text{FWE}_P$. We require that any procedure satisfy that the familywise error rate 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 only controls the FWE when all k null hypotheses are true is said to have weak control of the FWE. As remarked by Dudoit et al. (2002), 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,

$$c_{n,K}(\alpha, P) = \inf\{x : \ P\{\max_{j \in K} T_{n,j} \leq x\} \geq \alpha\} \ . \tag{1}$$

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

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

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

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

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, 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 (or not having to consider conditioning on the past). Because the Holm procedure (discussed later in Example 2.4) works in this way, one might hope that one can generally test the intersection hypothesis at any step without regard to hypotheses previously rejected. 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 2.1 (Idealized Stepdown Method)

1. Let $K_1 = \{1, \ldots, k\}$. If $T_{n,r_1} \leq c_{n,K_1}(1-\alpha)$, 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 $T_{n,r_2} \leq c_{n,K_2}(1-\alpha)$, then accept all remaining hypotheses and stop; otherwise, reject H_{r_2} and continue.

$\vdots$

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

$\vdots$

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

The above algorithm is an idealization for two reasons: the critical values may be impossible to compute and, 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 at least work approximately in large samples. First, we present an example to show that some condition is required to exhibit strong control.

Example 2.1 Suppose $T_{n,1}$ and $T_{n,2}$ are independent and normally distributed, with $T_{n,1} \sim N(\theta_1, (1+\theta_2)^{2p})$ and $T_{n,2} \sim N(\theta_2, (1+\theta_2)^{-2p})$, where $\theta_1 \geq 0$ and $\theta_2 \geq 0$. (The index n plays no role here, but we retain it for consistent notation.) Here, p is a suitable positive constant, chosen to be large. Also, let $\Phi(\cdot)$ denote the standard normal cumulative distribution function. The hypothesis H_i specifies $\theta_i = 0$ while H_i' specifies $\theta_i > 0$. Therefore, the first

step of Algorithm 2.1 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 i.i.d. $N(0,1)$. Specifically, accept both hypotheses if

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

otherwise, reject the hypothesis corresponding to the larger $T_{n,i}$. 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 $p \to \infty$. So, if $\alpha < 1/2$, for some large enough but fixed p, 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 which exhibits weak control but not strong control. (Single-step procedures are those where hypotheses are rejected on the basis of a single critical value; see Westfall and Young (1993).)

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

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

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

Theorem 2.1 *Let P denote the true distribution generating the data.*

(i) Assume for any K containing $I(P)$,

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

Then, the probability that Algorithm 2.1 rejects any $i \in I(P)$ is $\leq \alpha$; that is, $FWE_P \leq \alpha$.

(ii) Strong control persists if, in Algorithm 2.1, the critical constants $c_{n,K_j}(1-\alpha)$ are replaced by $d_{n,K_j}(1-\alpha)$ which satisfy

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

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

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

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

Remark 2.1 Under weak assumptions, one can show the sup over P of the probability that Algorithm 2.1 rejects any $i \in I(P)$ is equal to α. It then follows that the critical values cannot be made smaller, in hopes 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 one.)

Example 2.2 (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 which satisfies

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

for all P such that $I(P) \supset K$. Such a P_K may be referred to 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 $\theta_j \leq 0$, then intuition suggests 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 there exists a P_0 with $I(P_0) = \{1, \ldots, k\}$ such that the joint distribution of $\{T_{n,i} : i \in I(P_K)\}$ under P_K is the same as the distribution of $\{T_{n,i} : i \in I(P_K)\}$ under P_0. This condition says the (joint) distribution of the test statistics used for testing the hypotheses $H_i,\ i \in I(P_K)$ is unaffected by the truth or falsehood of the remaining hypotheses (and therefore we assume all hypotheses are true by calculating the distribution of the maximum under P_0). It follows that, in step j of Algorithm 2.1,

$$c_{n,K_j}(1-\alpha) = c_{n,K_j}(1-\alpha, P_{K_j}) = c_{n,K_j}(1-\alpha, P_0) = c_{n,K_j}(1-\alpha)\ ; \tag{9}$$

the outer equalities in (9) follow by the assumption (8) and the middle equality follows by the subset pivotality condition. Therefore, in Algorithm 2.1, 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 $\max(T_{n,i} : i \in K_j)$ 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 is considered in Finner and Roters (1998), and they recognize the importance of monotonicity for the validity of stepwise procedures. However, they also suppose 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 previously seen, such assumptions do not hold generally.

Example 2.3 To exhibit an example where condition (5) holds, but subset pivotality does not, suppose that $T_{n,1}$ and $T_{n,2}$ are independent, normally distributed, with $T_{n,1} \sim N(\theta_1, 1/(1+\theta_2^2))$ and $T_{n,2} \sim N(\theta_2, 1/(1+\theta_1^2))$. The hypothesis H_i specifies $\theta_i = 0$ while the alternative H_i' specifies $\theta_i > 0$. Then, it is easy to check that, with $K_1 = \{1, 2\}$,

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

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

Example 2.4 (The Holm Procedure) Suppose $-T_{n,i} \equiv \hat{p}_{n,i}$ is a p-value for testing H_i; that is, assume the distribution of $\hat{p}_{n,i}$ is Uniform on $(0,1)$ when H_i is true. Note that this assumption is much weaker than subset pivotality (if $k > 1$) because we are only making an assumption about the one-dimensional marginal distribution of the p-value statistic. Furthermore, we may assume the weaker condition

$$P\{\hat{p}_{n,i} \leq x\} \leq x$$

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

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

which is independent of P, and so

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

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 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 2.1(iii).

Part (iii) of Theorem 2.1 points toward a more general method that has strong control even when (5) is violated, and that can be much less conservative than the Holm procedure.

Corollary 2.1 *Let*

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

Then, if you replace $c_{n,K_j}(1-\alpha)$ by $c^*_{n,K_j}(1-\alpha)$ in Algorithm 2.1, strong control holds.

Corollary 2.1 is simply the closure principle of Marcus et al. (1976); also see Hommel (1986) and Theorem 4.1 of Hochberg and Tamhane (1987). Thus, in order to have a valid stepdown procedure, one must not only consider the critical value $c_{n,K}(1-\alpha)$ when testing an intersection hypothesis H_K, one must also compute all $c_{n,I}(1-\alpha)$ for $I \subset K$.

3 Random Critical Values and Randomization Tests

3.1 Preliminaries and a Basic Inequality

In general, the critical values used in Algorithm 2.1 are the smallest constants possible without violating the FWE. As a simple example, suppose X_i, $i = 1, \ldots, k$, are independent $N(\theta_i, 1)$, with the θ_i varying freely. The null hypothesis H_i specifies $\theta_i \leq 0$. Then,

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

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

$$P_0\{\max_{i\in K} X_i > c\} > P_0\{\max_{i\in K} X_i > c_{n,K}(1-\alpha)\} = \alpha \ .$$

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

However, the above only applies to nonrandom critical values and leaves open the possibility that critical values can be estimated, and therefore be random. That is, if we replace $c_{n,K}(1-\alpha)$ by some estimate $\hat{c}_{n,K}(1-\alpha)$, it can sometimes be smaller than $c_{n,K}(1-\alpha)$ as long as it is not with probability one. Of course, this is the typical case where critical values need to 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, and then the monotonicity assumptions of the previous section would then 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 directly lead to permutation tests.)

For the sake of increased generality, we instead proceed as follows. Suppose the $c_{n,K}(1-\alpha)$ in Algorithm 2.1 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 will assume two things. First, we will replace the monotonicity assumption (5) by montonicity of the estimated critical values; that is, for any $K \supset I(P)$,

$$\hat{c}_{n,K}(1-\alpha) \geq \hat{c}_{n,I(P)}(1-\alpha) \ . \tag{12}$$

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,

$$P\{\max(T_{n,i} : \ i \in I(P)) > \hat{c}_{n,I(P)}(1-\alpha)\} \leq \alpha \ . \tag{13}$$

We will show the basic inequality that the FWE_P is bounded above by left side of (13). So, if we can construct monotone critical values which also satisfy each one yields a level α for testing a single intersection hypothesis, then the next result says the stepdown procedure controls the FWE. Thus, the construction of a stepdown procedure is essentially 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 we can essentially enforce because they only depend on the data, can hold even if the corresponding nonrandom ones are not monotone.)

Theorem 3.1 *Let P denote the true distribution generating the data. Consider Algorithm 2.1 with $c_{n,K}(1-\alpha)$ replaced by estimates $\hat{c}_{n,K}(1-\alpha)$ satisfying (12).*

(i) Then,

$$FWE_P \leq P\{\max(T_{n,j} : \ j \in I(P)) > \hat{c}_{n,I(P)}(1-\alpha)\} \ . \tag{14}$$

(ii) Therefore, if the critical values also satisfy (13), then $FWE_P \leq \alpha$.

3.2 Permutation and Randomization Tests

Before applying Theorem 3.1, 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 Section 3.1.7 of Westfall and Young (1993)).

Based on data 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 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 will call the *randomization hypothesis*, allows for a general test construction.

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

As an example, consider testing the equality of distributions based on two independent samples $(Y_1, \ldots, Y_m)$ and $(Z_1, \ldots, Z_n)$. 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 $T(X)$ be any real-valued test statistic for testing H. Suppose the group $\mathbf{G}$ has M elements. Given $X = x$, let

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

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

$$m = M - [M\alpha] \ , \tag{15}$$

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

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

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

Under the randomization hypothesis, Hoeffding (1952) shows 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, the slightly conservative but *nonrandomized* test that rejects if $T(X) > T^m(X)$ is level α.

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

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

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

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

$$\hat{p} = \frac{1}{M} \sum_g I\{T(gX) \geq T(X)\} \ . \tag{16}$$

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

$$P\{\hat{p} \leq u\} \leq u \quad \text{ for all } 0 \leq u \leq 1 \ . \tag{17}$$

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

Because $\mathbf{G}$ may be large, one may resort to a stochastic approximation to construct the randomization test, for example, by randomly sampling transformations g from $\mathbf{G}$ with or without replacement. In the former case, for example, suppose $g_1, \ldots, g_{B-1}$ are i.i.d. and uniformly distributed on $\mathbf{G}$. Let

$$\tilde{p} = \frac{1}{B} \left[1 + \sum_{i=1}^{B-1} I\{T(g_i X) \geq T(X)\} \right] \ . \tag{18}$$

Then, it can be shown that, under the randomization hypothesis,

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

where this probability reflects variation in both X and the sampling of the g_i. Note that (19) holds for any B, and so the test that rejects when $\tilde{p} \leq \alpha$ is level α even when a stochastic approximation is employed. 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 (18) is based on the following simple fact.

Lemma 3.1 *Suppose $Y_1, \ldots, Y_B$ are exchangeable real-valued random variables; that is, their joint distribution is invariant under permutations. Let $\tilde{q}$ be defined by*

$$\tilde{q} = \frac{1}{B} \left[1 + \sum_{i=1}^{B-1} I\{Y_i \geq Y_B\} \right] \ .$$

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

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

$$T_{n,K} = \max(T_{n,i} : \ i \in K) \tag{20}$$

and reject H_K when

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

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 $I \subset K$,

$$\max(T_{n,i}(gX) : \ i \in K) \geq \max(T_{n,i}(gX) : \ i \in I) \ , \tag{21}$$

and so as g varies, the m_αth largest value of the left side of (21) is at least as large as the m_αth largest value of the right side.

Consequently, the critical values

$$\hat{c}_{n,K}(1-\alpha) = T_{n,K}^{(m_\alpha)} \ , \tag{22}$$

satisfy the monotonicity requirement of Theorem 3.1. 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 is true.

Corollary 3.1 *Suppose 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 (22) controls the FWE.*

Equivalently, in analogy with (16), we can compute p-values for testing H_K via

$$\hat{p}_{n,K} = \frac{1}{M} \sum_g I\{T_{n,K}(gX) \geq T_{n,K}(X)\} \ , \tag{23}$$

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 (18), randomly sample $g_1, \ldots, g_{B-1}$ from $\mathbf{G}$ and let

$$\tilde{p}_{n,K} = \frac{1}{B} \left[1 + \sum_{i=1}^{B-1} I\{T_{n,K}(g_i X) \geq T_{n,K}(X)\} \right] \ . \tag{24}$$

By an almost identical argument, we have the following.

Corollary 3.2 *Suppose the randomization hypothesis holds for a group* $\mathbf{G}$ *when testing any intersection hypothesis* H_K. *Consider the stepdown method which rejects* K_j *at stage* j *if* $\tilde{p}_{n,K_j} \leq \alpha$. *Then,* $FWE_P \leq \alpha$.

Remark 3.1 In the above corollaries, we have worked with the randomization construction using nonrandomized tests. A similar result would hold if we permit randomization.

Example 3.1 (Two Sample Problem With k Variables) Suppose $Y_1, \cdots, Y_{n_Y}$ is a sample of n_Y independent observations from a probability distribution P_Y and $Z_1, \cdots, Z_{n_Z}$ is a sample of n_Z observations from P_Z. Here, P_Y and P_Z are probability distributions on $\mathbf{R}^k$, with jth components denoted $P_{Y,j}$ and $P_{Z,j}$, respectively. The hypothesis H_j asserts $P_{Y,j} = P_{Z,j}$ and we wish to test these k hypotheses based on $X = (Y_1, \cdots, Y_{n_Y}, Z_1, \cdots, Z_{n_Z})$. Also, let $Y_{i,j}$ denote the jth component of Y_i and $Z_{i,j}$ denote the jth component of Z_i. As in Troendle (1995), we assume a semiparametric model. In particular, assume P_Y and P_Z are governed by a family of probability distributions Q_θ indexed by $\theta = (\theta_1, \ldots, \theta_k) \in \mathbf{R}^k$ (and assumed identifiable), so that P_Y has law $Q(\theta_Y)$ and P_Z has law $Q(\theta_Z)$. For concreteness, one may think of θ as being the mean vector, though this assumption is not necessary. Now, H_j can be viewed as testing $\theta_{Y,j} = \theta_{Z,j}$. 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 $n = n_Y + n_Z$, and for $x = (x_1, \cdots, x_n) \in \mathbf{R}^n$, let $gx \in \mathbf{R}^N$ be defined by $(x_{\pi(1)}, \cdots, x_{\pi(n)})$, where $(\pi(1), \cdots, \pi(n))$ is a permutation of $(1, 2, \cdots, n)$. Let $\mathbf{G}$ be the collection of all such g so that $M = n!$. Under the hypothesis $P_Y = P_Z$, gX and X have the same distribution for any g in $\mathbf{G}$.

Unfortunately, this $\mathbf{G}$ does not apply to any subset of the hypotheses. However, we just need a slight generalization to cover the example. Suppose that the test statistic $T_{n,j}$ used to test H_j only depends on the jth components of the observations, namely $Y_{i,j}$, $i = 1, \ldots, n_Y$ and $Z_{i,j}$, $i = 1, \ldots, n_Z$; this is a weak assumption indeed. In fact, let X_K be the data set consisting of the 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 data set, the randomization hypothesis holds. Specifically, under the null hypothesis $\theta_{Y,j} = \theta_{Z,j}$ for $j \in K$, X_K and gX_K have the same distribution (though X and gX need not). Also, for any $g \in \mathbf{G}$, $T_{n,j}(gX)$ and $T_{n,j}(X)$ have the same distribution under H_j, and similarly for any $K \subset \{1, \ldots, k\}$, $T_{n,K}(gX)$ and $T_{n,K}(X)$ 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 (21). Moreover, each intersection hypothesis can be tested by an exact level α randomization test (since inference for H_K is based only on X_K). Therefore, essentially the same argument leading to Corollaries 3.1 and 3.2 applies. In particular, even if we need to resort to approximate randomization tests at each stage, but 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) uses lengthy arguments to conclude only asymptotic control.

Remark 3.2 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 3.1, suppose

we are just interested in testing the hypothesis H'_j that the mean of $P_{Y,j}$ is the mean of $P_{Z,j}$ (assumed to exist). Then, the randomization test construction of this section fails because the randomization hypothesis need not hold. However, since the randomization procedure has monotone critical values (as this is only a property of how the data is used), Theorem 3.1(i) 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 is studied in 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 employed; the comparison of randomization and bootstrap tests has been studied in Romano (1989) and it is shown they are often quite close, at least when the randomization hypothesis holds.

Example 3.2 (Problem of Multiple Treatments) Consider the one-way anova model. We are given $k+1$ independent samples, with the jth sample having n_j i.i.d. observations $X_{i,j}$, $i = 1, \ldots, n_j$. Suppose $X_{i,j}$ has distribution P_j. The problem is to test the hypotheses of k treatments with a control; that is, $H_i : \ P_i = P_{k+1}$. (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, this group is not valid. A simple remedy is to permute only within subsets; that is, to test any subset hypothesis H_K, only consider those permutations that permute observations within the sample $X_{i,k+1}$ and the samples $X_{i,j}$ with $j \in K$. Therefore, one computes 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 analogue of Corollary 2.1, if one is willing to compute critical values for all subset hypotheses; that is, replace $\hat{c}_{n,K_j}(1-\alpha)$ by

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

On the other hand, this can be computationally prohibitive. Such issues were raised by Petrondas and Gabriel (1983) (although the problem was not framed in terms of a montonicity requirement). Using the critical value $\hat{c}^*_{n,K_j}(1-\alpha)$ is based on the closure principle of Marcus et al. (1976) and is also similar to (2.13) of Westfall and Young (1993). However, we will shortly see that the lack of monotonicity of critical values is only a finite sample concern; see Example 4.2.

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 2.1 fails. The assumptions are identical to the weakest assumptions available for the contruction

of asymptotically valid tests of a single hypothesis, which are used in many resampling schemes, and so one cannot expect to improve them without improving the now well-developed theory of resampling methods for testing a single hypothesis.

Of course, Corollary 2.1 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 is computed by a bootstrap or resampling method, the number of computations could get quite large for even moderate k. Not only will we provide weak conditions, but we will consider a method that only requires *one* set of bootstrap resamples, as well as a method based on *one* set of subsamples.

In order to accomplish this without having to invoke an assumption like subset pivotality, we will consider resampling schemes that do *not* obey the constraints of the null hypothesis. Such schemes, as discussed in 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 is even advocated as a principle in Hall and Wilson (1991), and it is enforced throughout Westfall and Young (1993). While 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 you can construct good or valid confidence intervals, then you can construct good or valid tests, and conversely. But, there is no dispute that resampling the empirical distribution to construct a confidence interval for a single parameter can indeed produce very desirable intervals, which 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 implausible that one resampling scheme obeying the constraints of all hypotheses would work 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 sheme for testing each subset hypothesis. Second, even if one knew how to resample at each stage, there is increased computation. Our approach avoids these complications.

Before embarking on the general theory, a motivating example is presented to fix ideas.

Example 4.1 (Testing Correlations) Suppose $X_1, \ldots, X_n$ are i.i.d. random vectors in $\mathbb{R}^s$, so that $X_i = (X_{i,1}, \ldots, X_{i,s})$. Assume $E|X_{i,j}|^2 < \infty$ and $Var(X_{i,j}) > 0$, 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 now by 2 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) Aitken (1971), if $s = 3$, $H_{1,2}$ and $H_{1,3}$ are true but $H_{2,3}$ is false, the joint

limiting distribution of $n^{1/2}(T_{n,1,2}, T_{n,1,3})$ is bivariate normal with means zero, variances one, and correlation $\rho_{2,3}$. As acknowledged by Westfall and Young (1993), 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 $T_{n,i}$, $i \in K$ under P, with corresponding joint c.d.f. $G_{n,K}(x, P)$, $x \in \mathbb{R}^{|K|}$. Also, let $H_{n,K}(P)$ denote the distribution of $\max\{T_{n,i} : i \in K\}$ under P. As in the previous section, its $1-\alpha$ quantile is denoted $c_{n,K}(1-\alpha, P)$. Also, the symbols $\xrightarrow{L}$ and $\xrightarrow{P}$ will denote convergence in law (or distribution) and convergence in probability, respectively.

Typically, the asymptotic behavior of $G_{n,I(P)}(P)$ is governed by one of the following two possibilities. Either it has a nondegenerate limiting distribution, or it 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 we are testing a population mean $\mu(P)$ is ≤ 0 versus > 0 based on an i.i.d. 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 $T_n \xrightarrow{L} N(0, \sigma^2(P))$. 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 $Z \sim N(0,1)$. But, under $\mu(P) < 0$, T_n' converges in probability to 0. Note, 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 will distinguish between the following assumptions, which are only imposed when $K = I(P)$ is the set of true hypotheses.

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

$$G_{n,I(P)}(P) \xrightarrow{L} G_{I(P)}(P) \ . \tag{25}$$

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

$$H_{I(P)}(x, P) \quad \text{is continuous and strictly increasing at } x = c_{I(P)}(1-\alpha, P) \ . \tag{26}$$

Note that the continuity condition in (26) is satisfied if the $|I(P)|$ univariate marginal distributions of $J_{I(P)}(P)$ are continuous. Also, the strictly increasing assumption can be weakened as well, but it holds in all known examples where the continuity assumption holds, as typical limit distributions are of the Gaussian, Chi-squared, etc. type. Actually, the strictly increasing assumption can be removed entirely (see Remark 1.2.1 of Politis et al. (1999)).

Assumption A2 Under P, $G_{n,I(P)}(P)$ converges weakly to a point mass at $d = d(P)$, where $d = (d_1(P), \ldots, d_{|I(P)|}(P))$ is a vector of $|I(P)|$ components. (In the case where $d_i(P) = -\infty$, we mean $T_{n,i}$ 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 2.1 with $c_{n,K}(1-\alpha)$ replaced by some estimates $\hat{c}_{n,K}(1-\alpha)$. We will consider some concrete choices later.

Theorem 4.1 *(i) Fix P and suppose Assumption A1 holds, so that (25) and (26) hold. Assume 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 $\epsilon > 0$*

$$\hat{c}_{n,K}(1-\alpha) \geq c_{I(P)}(1-\alpha) - \epsilon \quad \text{with probability } \to 1 \ . \tag{27}$$

Then, $\limsup_n FWE_P \leq \alpha$.

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

$$\hat{c}_{n,K}(1-\alpha) \geq \hat{c}_{n,I}(1-\alpha) \quad \text{whenever } I \subset K \ . \tag{28}$$

Then, (27) holds for all $K \supset I(P)$ if it holds in the special case $K = I(P)$. Therefore, if Assumption A1 and the montonicity condition (28) hold, and

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

then $\limsup_n FWE_P \leq \alpha$.

(iii) Fix P and suppose Assumption A2 holds. Also, assume the monotonicity condition (28). If, for some $\epsilon > 0$,

$$\hat{c}_{n,I(P)}(1-\alpha) > \max\{d_i(P) : \ i \in I(P)\} + \epsilon \quad \text{with probability } \to 1, \tag{30}$$

then $\limsup_n FWE_P = 0$.

Note that Assumption A1 implies

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

In part (i) of Theorem 4.1, we replace the monotonicity requirement of Theorem 3.1 by a weak asymptotic monotonicty requirement (27).

In general, the point of Theorem 4.1 is that $\limsup_n \text{FWE}_P \leq \alpha$ regardless of whether the convergence of the null hypotheses satisifies Assumption A1 or Assumption A2, at least under reasonable behavior of the estimated critical values. Moreover, the monotonicity condition (28) assumed in parts (ii) and (iii) will be shown to 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 (29); 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 controlloing the FWE is reduced to the problem of controlling the Type 1 error of a single test. This will be further clarified for specific choices of estimates of the critical values.

Before applying Theorem 4.1 (ii), (iii), which assumes monotonicity of critical values, we demonstrate consistency without the assumption of monotonicity. In this regard, a simple alternative to Theorem 4.1 (i) is the following.

Theorem 4.2 *Fix P and suppose Assumption A1 holds. Suppose 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 one. 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,* $\limsup_n FWE \leq \alpha$.

Example 4.2 (Example 3.2, revisited) In the setup of Example 3.2, suppose 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 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 $n_i/n \to \lambda_i \in (0,1)$. Let $\mu(P_i)$ denote the true mean of P_i, assumed to exist; also assume the variance of P_i is finite. Then, Theorem 4.2 applies to any P for which, if $j \notin I(P)$, $\mu(P_i) \neq \mu(P_{k+1})$ (which, of course, is not the same as $P_i \neq P_{k+1}$). Indeed, $T_{n,i} \to \infty$ in probability. Also, using arguments as in 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, 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 will develop a concrete construction based on the bootstrap. For now, we suppose hypothesis H_i is specified by $\{P : \ \theta_i(P) \leq 0\}$ for some real-valued parameter θ_i. Suppose $\hat{\theta}_{n,i}$ is an estimate of θ_i. Also, let $T_{n,i} = \tau_n \hat{\theta}_{n,i}$ for some nonnegative (nonrandom) sequence $\tau_n \to \infty$. The sequence τ_n is introduced for asymptotic purposes so that a limiting distribution for $\tau_n \hat{\theta}_{n,i}$ exists when $\theta_i(P) = 0$.

Remark 4.1 Typically, $\tau_n = n^{1/2}$. Also, it is possible to let τ_n vary with the hypothesis i. Extensions to cases where τ_n depends on P are also possible, using ideas in Bertail et al. (1999).

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

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

Let $L_{n,K}(P)$ denote the distribution under P of $\max\{\tau_n[\hat{\theta}_{n,i} - \theta_i(P)] : i \in K\}$, 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 $L_{n,K}(P)$ has a limiting distribution $L_K(P)$.

We will further assume

Assumption B1(ii) $L_K(P)$ is continuous and strictly increasing on its support.

Under Assumption B1, it follows that

$$b_{n,K}(1 - \alpha, P) \to b_K(1 - \alpha, P) , \tag{31}$$

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

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

Let $\hat{Q}_n$ be some estimate of P. Then, a nominal $1 - \alpha$ level bootstrap confidence region for the subset of parameters $\{\theta_i(P) : i \in K\}$ is given by

$$\{(\theta_i : i \in K) : \max_{i \in K} \tau_n[\hat{\theta}_{n,i} - \theta_i] \leq b_{n,K}(1 - \alpha, \hat{Q}_n)\}$$

$$= \{(\theta_i : i \in K) : \theta_i \geq \hat{\theta}_{n,i} - \tau_n^{-1} b_{n,K}(1 - \alpha, \hat{Q}_n)\} .$$

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

$$\hat{c}_{n,K}(1 - \alpha) = b_{n,K}(1 - \alpha, \hat{Q}_n) , \tag{32}$$

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

Note that, regardless of asymptotic behavior, the monotonicity assumption (28) is always satisfied for the choice (32). 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, while $b_{n,K}(1 - \alpha, Q)$ is the $1 - \alpha$ quantile of these same $|I|$ variables together with $|K| - |I|$ variables.

Therefore, in order to apply Theorem 4.1 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 $K = I(P)$. For

this, we further assume the usual conditions for bootstrap consistency when testing the *single* hypothesis that $\theta_i(P) \leq 0$ for all $i \in I(P)$; that is, we assume the bootstrap consistently estimates the joint distribution of $\tau_n[\hat{\theta}_{n,i} - \theta_i(P)]$ for $i \in I(P)$. Specifically, consider the following.

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

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

Theorem 4.3 *Fix P satisfying assumption B1. Let $\hat{Q}_n$ be an estimate of P satisfying B2. Consider the stepdown method in Algorithm 2.1 with $c_{n,K}(1-\alpha)$ replaced by $b_{n,K}(1-\alpha, \hat{Q}_n)$. Then,* $\limsup_n FWE_P \leq \alpha$.

Example 4.3 (Continuation of Example 4.1) The analysis of sample correlations is a special case of the smooth function model studied in Hall (1992), and the bootstrap approach is valid for such models.

Remark 4.2 The above analysis extends to the two-sided case. Simply change assumption B1(ii) to reflect the distribution of $\max\{\tau_n|\hat{\theta}_{n,i} - \theta_i(P)| : \ i \in K\}$, and the theorem holds.

Remark 4.3 The main reason why the bootstrap works here can be traced to the simple result Theorem 3.1. The bootstrap approach, by resampling from a fixed distribution, generates monotone critical values. Therefore, since 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 also can deduce $\limsup_n \text{FWE}_P \leq \alpha + O(\epsilon_n)$, so that efficient bootstrap methods for single tests then translate into efficient bootstrap methods for multiple tests.

Remark 4.4 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 one. However, one can also study the behavior of our procedures against contiguous alternatives so that not all false hypotheses are rejected with probability tending to one 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 will be bounded by α.

Remark 4.5 The construction developed in this subsection can be extended to the case of studentized test statistics. The details are straightforward and left to the reader.

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 i.i.d. sample $X_1, \ldots, X_n$ from P, and $T_{n,i} = T_{n,i}(X_1, \ldots, X_n)$ is the test statistic we wish to use for testing H_i. To describe the test construction, fix a positive integer $b \leq n$ let $Y_1, \ldots, Y_{N_n}$ be equal to the $N_n = \binom{n}{b}$ subsets of $\{X_1, \ldots, X_n\}$, ordered in any fashion. Let $T_{b,i}^{(j)}$ be equal to the statistic $T_{b,i}$ evaluated at the data set Y_j. Then, for any subset $K \subset \{1, \ldots, k\}$, the joint distribution of $(T_{n,i} : i \in K)$ can be approximated by the empirical distribution of the $\binom{n}{b}$ values $(T_{b,i}^{(j)} : \ i \in K)$. In other words, for $x \in \mathbf{R}^k$, the true joint c.d.f. of the test statistics evaluated at x,

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

is estimated by the subsampling distribution

$$\hat{G}_{n,\{1,\ldots,k\}}(x) = \binom{n}{b}^{-1} \sum_j I\{T_{b,1}^{(j)} \leq x_1, \ldots, T_{b,k}^{(j)} \leq x_k\} \ . \tag{33}$$

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 (33) on that subset of variables. So, $\hat{G}_{n,K}$ refers to the empirical distribution of the values $(T_{n,i}^{(j)} : \ i \in K)$. (In essence, one only has to 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 $\max(T_{n,i}^{(j)} : \ i \in K)$; that is,

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

Also, let

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

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

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

$$\hat{c}_{n,K}(1-\alpha) \geq \hat{c}_{n,I}(1-\alpha) \ ; \tag{34}$$

and so the monotonicity assumption in Theorem 4.1 holds (and 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 4.4 *Suppose Assumption A1 holds. Let* $b/n \to 0$ *and* $b \to \infty$.

(i). The subsampling approximation satisfies

$$\rho\left(\hat{G}_{n,I(P)}, G_{n,I(P)}(P)\right) \xrightarrow{P} 0 \ . \tag{35}$$

(ii) The subsampling critical values satisfy

$$\hat{c}_{n,I(P)}(1-\alpha) \xrightarrow{P} c_{I(P)}(1-\alpha) \ . \tag{36}$$

(iii). Therefore, using Algorithm 2.1 with $c_{n,K}(1-\alpha, P)$ *replaced by the estimates* $\hat{c}_{n,K}(1-\alpha)$ *results in* $\limsup_n FWE \leq \alpha$.

Example 4.4 (Cube root asymptotics) Kim and Pollard (1990) show that a general class of M-estimators converge at rate $\tau_n = n^{1/3}$ to a non-normal limiting distribution. As result, inconsistency of the bootstrap typically follows. Rodríguez-Poo et al. (2001) demonstrate 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 4.4 in the context of cube root asymptotics can be established.

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

5 Simulation Study

This section presents a small simulation study in the context of Example 4.1. We generate random vectors $X_1, \ldots, X_{100}$ from a 10-dimensional multivariate normal distribution. Hence, there are a total of $k = \binom{10}{2} = 45$ pairwise correlations to test. Each individual null hypothesis is $H_{i,j}$: $\rho_{i,j} = 0$; and each individual alternative hypothesis is two-sided. We apply the stepdown bootstrap construction of Subsection 4.2, resampling from the empirical distribution. As a special case, we also look at the single-step method based on $K = \{1, \ldots, k\}$ only. The nominal FWE levels are $\alpha = 0.05$ and $\alpha = 0.1$. 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 0.3, for $j = 2, \ldots, 10$, and the remaining correlations are equal to 0. In the third scenario, all correlations are equal to 0.3.

Table 1 reports the results based on 5,000 repetitions. The number of bootstrap resamples is $B = 500$ always. The results demonstrate the good control of the FWE in finite sample and the increased power of the stepdown method compared to the single-step method.

6 Empirical Application

Westfall and Young (1993, Example 6.4) apply a multiple testing method for 10 pairwise correlations. Each individual null hypothesis is that corresponding pairwise population correlation is equal to zero; and each individual alternative hypothesis is two-sided. The reader is referred to their Example 6.4 for the details of the real data set. Westfall and Young (1993) carry out a bootstrap multiple test under the assumption of complete independence. As they admit, this is a conservative approach in general. Instead we apply the stepdown bootstrap construction of Subsection 4.2, resampling from the empirical distribution.

Table 2 compares the adjusted P-values of Westfall and Young (1993) to ours. The conservativeness of the Westfall and Young (1993) method can be clearly appreciated.

A Proofs

Proof of Theorem 2.1

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

$$T_{n,r_{\hat{j}}} > c_{n,K_{\hat{j}}}(1-\alpha) \ . \tag{37}$$

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

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

and so

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

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

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

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

$$\max(T_{n,j},\ j \in I(P)) > c_{n,I(P)}(1-\alpha,P) \ . \tag{38}$$

Therefore, the probability of a Type 1 error is bounded above by the probability of the event (38), which be definition has probability bounded above by α. The proof of (ii) is obvious because the procedure becomes more conservative. The proof of (iii) holds by the proof of (i) upon replacing the constants $c_{n,K_{\hat{j}}}(1-\alpha)$ by $d_{n,K_{\hat{j}}}(1-\alpha)$. ■

Proof of Corollary 2.1

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 2.1 (ii) and (iii). Clearly,

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

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

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

and so (7) holds. ■

Proof of Theorem 3.1

As in the argument of Theorem 2.1, the event a false rejection occurs is the event

$$\max\{T_{n,j}:\ j \in I(P)\} > \hat{c}_{n,K_{\hat{j}}}(1-\alpha)\ , \tag{39}$$

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

$$\hat{c}_{n,K_{\hat{j}}}(1-\alpha) \geq \hat{c}_{n,I(P)}(1-\alpha) \tag{40}$$

and so (i) follows. Part (ii) follows immediately from (i). ■

Proof of Theorem 4.1

As in the proofs of Theorems 2.1 and 3.1, namely (39), it suffices to show

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

But assumption (27) implies

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

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

$$\limsup_n FWE_P \leq P\{\max(T_j,\ j \in I(P)) > c_{I(P)}(1-\alpha) - \epsilon\}\ ,$$

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 (ii), since (27) holds when $K = I(P)$, then it must hold for any K containing $I(P)$, by assumption (28).

To prove (iii), the probability of false rejection, i.e. the event (39), 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 (30). ■

Proof of Theorem 4.2

Following the proof of Theorem 4.1 (i), the random index $\hat{j}$ is equal to $k - |I(P)| + 1$ with probability tending to one, and this index is no longer random; that is, with probability tending to one, we first reject all false hypotheses and then commit a false rejection when we get to the stage where we are testing the $|I(P)|$ true hypotheses. But then, Assumption A1 allows us to conclude control of the FWE. ■

Proof of Theorem 4.3

Fix P and assume $\theta_i(P) = 0$ for at least one $i \in I(P)$. Then, by the comments leading up to the statement of the theorem, the conditions of Theorem 4.1 (ii) are satisfied if we can verify

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

But by the Continuous Mapping Theorem, the assumption B2 implies

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

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

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

and

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

So it suffices to show

$$\liminf b_{n,I(P)}(1-\alpha, P) \to c_{I(P)}(1-\alpha, P) \ . \tag{41}$$

But, for $\theta_i(P) \leq 0$,

$$\tau_n[\hat{\theta}_{n,i} - \theta_i(P)] \geq \tau_n \hat{\theta}_{n,i} = T_n \ ,$$

which implies

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

But, the right term converges to $c_{I(P)}(1-\alpha, P)$, and so (41) follows.

Next, assume P has $\theta_i(P) < 0$ for all $i \in I(P)$. Then, we just need to verify the conditions of Theorem 4.1 (iii). All that is left to verify is, for some $\epsilon > 0$,

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

with probability tending to one. But, the right side here is $-\infty$ (for any finite ϵ), so it just suffices to verify the left side is $O_P(1)$. But, by B2, it suffices to show $b_{n,I(P)}(1-\alpha, P)$ is bounded away from $-\infty$, which follows by (31). ■

Proof of Theorem 4.4

The proof of (i) is the essential subsampling argument, which derives from (33) being a U-statistic; see Politis et al. (1999), Theorem 2.6.1, where one statistic is treated, but the argument is extendable to the simultaneous estimation of the joint distribution. The result (ii) follows as well. To verify (iii), apply Theorem 4.1 (ii). The monotonicity requirement follows by (34) and (29) follows by (ii). ■

References

Aitken, M. (1969). Some tests for correlation matrices. *Biometrika*, 56:443–446.

Aitken, M. (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 Series B*, 57(1):289–300.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, 29(4):1165–1188.

Beran, R. (1986). Simulated power functions. *Annals of Statistics*, 14:151–173.

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. (2002). Multiple hypothesis testing in microarray experiments. Technical report, Division of Biostatistics, U.C. Berkeley. Available at `http://www.bepress.com/ucbbiostat/paper110/`.

Finner, H. and Roters, M. (1998). Asymptotic comparison of step-down and step-up multiple test procedures based on exchangeable test statistics. *Annals of Statistics*, 26:505–524.

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

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*. Wiley, New York.

Hoeffding, W. (1952). The large-sample power of tests based on permutations of observations. *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. *Annals of Statistics*, 18:191–219.

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(384):949–957.

Politis, D. N., Romano, J. P., and Wolf, M. (1999). *Subsampling*. Springer, New York.

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.

Romano, J. (1988). A bootstrap revival of some nonparametric distance tests. *Journal of the American Statistical Association*, 83(403):698–708.

Romano, J. (1989). Bootstrap and randomization tests of some nonparametric hypotheses. *Annals of Statistics*, 17:141–159.

Romano, J. (1990). On the behavior of randomization tests without a group invariance assumption. *Journal of the American Statistical Association*, 85(411):686–692.

Troendle, J. (1995). A stepwise resampling method of multiple hypothesis testing. *Journal of the American Statistical Association*, 90:370–378.

Westfall, P. H. and Young, S. S. (1993). *Resampling-Based Multiple Testing: Examples and Methods for P-Value Adjustment*. John Wiley, New York.

B Tables

Table 1: Empirical FWEs and average number of false hypotheses rejected for both the single-step construction and general stepdown construction of Subsection 4.2. The nominal levels are $\alpha = 5\%$ and $\alpha = 10\%$. Observations are i.i.d. multivariate normal, the number of observations is $n = 100$, and the number of pairwise correlations is $k = 45$. The number of repetitions is 5,000 per scenario and the number of bootstrap resamples is B = 500.

All $\rho_{i,j} = 0$				
Level α	FWE (single-step)	FWE (stepdown)	Rejected (single-step)	Rejected (stepdown)
5	4.6	4.6	0.0	0.0
10	9.8	9.8	0.0	0.0
All $\rho_{1,j} = 0.3$ and remaining $\rho_{i,j} = 0$				
Level α	FWE (single-step)	FWE (stepdown)	Rejected (single-step)	Rejected (stepdown)
5	4.0	4.2	3.7	3.8
10	8.5	8.8	4.5	4.6
All $\rho_{i,j} = 0.3$				
Level α	FWE (single-step)	FWE (stepdown)	Rejected (single-step)	Rejected (stepdown)
5	0.0	0.0	21.1	25.4
10	0.0	0.0	26.4	30.9

Table 2: Sample correlations and P-values for the data of Example 6.4 of Westfall and Young (1993). 'W-Y P-value' denotes the adjusted P-value of Westfall and Young; 'Step P-value' denotes the adjusted bootstrap P-value of Subsection 4.2 (based on $B = 5,000$ bootstrap resamples).

Variables	Sample correlation	Raw P-value	W-Y P-value	Step P-value
(SATdev, % Black)	-0.5089	.0002	.0019	.0016
(Salary, Crime)	0.4902	.0003	.0030	.0028
(% Black, Crime)	0.4844	.0004	.0036	.0034
(SATdev, S/T Ratio)	-0.3864	.0061	.0404	.0346
(SATdev, Crime)	-0.3033	.0341	.1843	.1483
(S/T Ratio, Crime)	0.2290	.1135	.4485	.3921
(S/T Ratio, % Black)	0.1732	.2341	.6474	.5986
(SATdev, Salary)	0.0980	.5030	.8753	.8572
(Salary, % Black)	-0.0354	.8090	.9641	.9645
(S/T Ratio, Salary)	0.0045	.9754	.9759	.9761