Published in final edited form as:
Biom J. 2008 October ; 50(5): 745–755. doi:10.1002/bimj.200710456.

Multiple Testing with Minimal Assumptions

Peter H. Westfall[*,1] and **James F. Troendle**[2]

[1] Texas Tech University, Lubbock TX USA

[2] National Institutes of Health, Bld. 6100, Room 7B05, Bethesda, MD 20892, USA

Summary

Resampling-based multiple testing methods that control the Familywise Error Rate in the strong sense are presented. It is shown that no assumptions whatsoever on the data-generating process are required to obtain a reasonably powerful and flexible class of multiple testing procedures. Improvements are obtained with mild assumptions. The methods are applicable to gene expression data in particular, but more generally to any multivariate, multiple group data that may be character or numeric. The role of the disputed "subset pivotality" condition is clarified.

Keywords

Bootstrap; Exchangeability; Permutation; Resampling; Subset pivotality

1 Introduction

With the recent "-omics" revolution, there is great interest in high-dimensional multiple testing, where the number of variables far exceeds the sample size. Gene expression is a prototype application, but the applications are much broader. "Resampling" is a general term that encompasses bootstrap, permutation, and parametric simulation-based analyses; "resampling-based multiple testing" refers to the use of such methods in multiple testing applications. Resampling methods have become popular for "-omics" because they (a) require fewer assumptions (e.g. normality) about the data-generating process, thereby yielding procedures that are more robust, (b) utilize data-based distributional characteristics, e.g. discreteness and correlation structure, to make tests more powerful, and (c) scale up reasonably well to high-dimensional settings, particularly with modern computing.

A lot has been made of the "subset pivotality condition" coined by Westfall and Young (1993) for resampling-based multiple tests. It has been portrayed in the literature as too stringent; for example, Romano, Shaikh, and Wolf (2007) state

> "the … condition of subset pivotality … assumed in … Westfall and Young (1993) … is quite restrictive."

Our purposes are (a) to clarify the role of subset pivotality, (b) to show that it is hardly restrictive, and (c) to clarify what is actually needed for validity of multiple testing. We will show that resampling-based multiple testing procedures can be valid, powerful, and control the familywise error rate (FWE) in the strong sense (Hochberg and Tamhane, 1987, define

[*] Corresponding author: peter.westfall@ttu.edu, Phone: +001 806 742 2174, Fax: +01 806 742 2174.

Conflict of Interests Statement
The authors have declared no conflict of interest.

strong control of the FWE), with no assumptions whatsoever on the data-generating process, yet where subset pivotality holds nevertheless. Slightly more power is available if one is willing to make one simple assumption about the data-generating process, an assumption distinct from, yet which is often confused with, subset pivotality.

Section 2 clarifies the role of the subset pivotality condition in multiple testing. Section 3 dispenses with assumptions altogether, notes that subset pivotality holds, and shows that valid, powerful, flexible FWE-controlling procedures are available. Section 4 shows how tests from Section 3 can be made more powerful, using a simple assumption that is not directly related to subset pivotality. Concluding remarks are given in Section 5.

2 The Role of Subset Pivotality in Closed Testing

Before describing the subset pivotality condition and its purpose, it is necessary to discuss the issues of choice of test statistic and computations. The reason is that the subset pivotality condition is only needed to simplify computations for resampling-based closed testing procedures.

2.1 Resampling-based multiple testing and closure

The closure principle of Marcus Peritz and Gabriel (1976) provides a unifying theory for hypothesis testing to control the FWE in the strong sense. (Hereafter “FWE control” is always assumed to mean “in the strong sense.”) Denoting the null hypotheses H_i, $i = 1; \dots ; m$, a closed testing procedure (CTP) requires that all intersection hypotheses be tested. Define intersection hypotheses $H_I = \cap_{i \in I} H_i$, for $I \subseteq S :=\{1; \dots ; m\}$. Let $\mathcal{H} = \{H_I \mid I \subseteq S\}$ denote of the set distinct intersection hypotheses. Then a CTP is one in which $H \in \mathcal{H}$ is rejected iff g H_- is rejected for every $H_- \in \mathcal{H}$ such that $H_- \subseteq H$. When the tests of H_{I} are α-level (not multiplicity adjusted) tests, the CTP controls the FWE at level α

For the purposes of this paper, a “resampling-based multiple testing procedure” is defined as one where each intersection H_I is tested using a resampling-based test. Applications of resampling not using closure exist, but the subset pivotality condition is best understood in the context of closure.

2.2 Choice of test statistic

Closed testing is very flexible in that any test statistic may be used to test the intersections H_I. There are many choices, including *F*- and related tests, Fisher combination tests, O'Brien-type tests, Simes-type tests, and weighted variants of all these tests. The choice of test statistic to use should primarily be based on power considerations. Once a powerful test statistic is chosen, resampling can be used to ensure that the test is robust to violations of distributional and/or dependence assumptions. An example appears in Dmitrienko, Offen and Westfall (2006), who show how to bootstrap the Simes test parametrically in a closed testing framework to accommodate correlation structure.

2.3 Computational issues: the MaxT test and subset pivotality condition

While power is the main concern for choice of a test statistic, expediency becomes important when m is large. There are $O(2^m)$ intersection hypotheses H_I, and if m is large, it is computationally impossible to test every single H_I. However, the computational burden can be eased dramatically if one is willing to

A: test each hypotheses H_I using a “Max” statistic $\max_{i \in I} T_i$, possibly sacrificing power, and

B: assume a model that implies "subset pivotality", which states that the distributions of $\max_{i\in I} T_i \mid H_I$ and $\max_{i\in I} T_i \mid H_{\{1;\ldots;m\}}$ are identical, for all $I \subset \{1; \ldots, m\}$.

If A and B are adopted, one need only test m hypotheses corresponding to the ordered t_i rather than all 2^m intersections; further, resampling can be done simultaneously under a global null $H\{1;\ldots,m\}$, rather than separately for each intersection. Note that "Max" subsumes "Min", where the test statistic is $-T_i$; the Min P test is commonly used (Westfall and Young, 1993).

To illustrate, suppose the observed test statistics are $t_1 \geq \ldots \geq t_m$, corresponding to hypotheses $H_1; \ldots,H_m$ (ordered in this way without loss of generality), and that larger t_i suggest alternative hypotheses. Suppose a p-value for testing H_I using the statistic $\max_{i\in I} T_i$ is available, then

$$p_I = P\left(\max_{i\in I} T_i \geq \max_{i\in I} t_i | H_I\right),$$

and H_I is rejected at unadjusted level α if $p_I \leq \alpha$. Applying closure, A, and B, we have the following algorithm for rejecting $H_1, H_2, \ldots$ in sequence using what we call the "Main Algorithm."

Main Algorithm:

1. By closure,

$$\text{reject } H_1 \text{ if} \max_{I:I\supseteq\{1\}} P\left(\max_{i\in I} > T_i \geq \max_{i\in I} t_i | H_I\right) \leq \alpha.$$

But if I ⊇ {1}, then $\max_{\mathrm{i}\in\mathrm{I}} t_i = t_1$, hence the rule is

$$\text{reject } H_1 \text{ if} \max_{I:I\supseteq\{1\}} P\left(\max_{i\in I} T_i \geq t_1 | H_I\right) \leq \alpha.$$

Using subset pivotality (B), the rule becomes

$$\text{reject } H_1 \text{ if} \max_{I:I\supseteq\{1\}} P\left(\max_{i\in I} T_i \geq t_1 | H_{\{1,\ldots,m\}}\right) \leq \alpha.$$

Use of the "Max" statistic (A) implies

$$P\left(\max_{i\in I} T_i \geq t_1 | H_{\{1,\ldots,m\}}\right) \leq P\left(\max_{i\in J} T_i \geq t_1 | H_{\{1,\ldots,m\}}\right) \text{ for } \quad I \subseteq J.$$

Hence, by subset pivotality and by use of the "Max" statistic, the rule by which we reject H_1 simplifies to this:

$$\text{reject } H_1 \text{ if } P\left(\max_{i\in\{1,\ldots,m\}} T_i \geq t_1 | H_{\{1,\ldots,m\}}\right) \leq \alpha.$$

2. Again by closure and subset pivotality,

$$\text{reject } H_2 \text{ if} \max_{I:I\supseteq\{2\}} P\left(\max_{i\in I} T_i \geq \max_{i\in I} t_i | H_{\{1,\ldots,m\}}\right) \leq \alpha.$$

If $I \supseteq \{1\}$, then $\max_{i\in I} t_i = t_1$; else $\max_{i\in I} t_i = t_2$. Partitioning the set {I : I ⊇ {2}} into two sets,

$$\boldsymbol{S}_1 = \{I{:}I \supseteq \{1,2\}\}, \qquad \boldsymbol{S}_2 = \{I{:}I \supseteq \{2\}, 1 \notin I\},$$

we require

$$P\left(\max_{i\in I} T_i \geq t_1 | H_{\{1,\ldots,m\}}\right) \leq \alpha \quad \text{for all} \quad I \in \boldsymbol{S}_1$$

and

$$P\left(\max_{i\in I} T_i \geq t_2 | H_{\{1,\ldots,m\}}\right) \leq \alpha \quad \text{for all} \quad I \in \boldsymbol{S}_2.$$

Since we are using the “Max” statistic, these conditions are equivalent to the following rejection rule: reject H_2 if

$$P\left(\max_{i\in\{1,\ldots,m\}} T_i \geq t_1 | H_{\{1,\ldots,m\}}\right) \leq \alpha$$

and

$$P\left(\max_{i\in\{2,\ldots,m\}} T_i \geq t_2 | H_{\{1,\ldots,m\}}\right) \leq \alpha.$$

j.: Continuing in this fashion, the rule is reject H_j if

$$P\left(\max_{i\in\{1,\ldots,m\}} T_i \geq t_1 | H_{\{1,\ldots,m\}}\right) \leq \alpha$$

and

$$P\left(\max_{i\in\{2,\ldots,m\}} T_i \geq t_2 | H_{\{1,\ldots,m\}}\right) \leq \alpha$$

and … and

$$P\left(\max_{i\in\{j,\ldots,m\}} T_i \geq t_j | \, !H_{\{1,\ldots,m\}}\right) \leq \alpha.$$

One need not use resampling at all to apply the method. It is only called a “resampling-based” procedure if one uses resampling to obtain the probabilities $P(\max_{i\in\{j,\ldots,m)} T_i \geq t_j \mid H\{1,\ldots,m)$.

At step j the rule is equivalently stated in terms of p-values for the composite hypotheses as

$$\text{reject } H_j \text{ if } \max_{i\leq j} p_{\{i,\ldots,m\}} \leq \alpha;$$

hence the rule reduces to

$$\text{reject } H_j \text{ if } \tilde{p}_j \leq \alpha,$$

where

$$\tilde{p}_j := \max_{i\leq j} p\{i,\ldots,m\}$$

is called the “adjusted p-value” (Westfall and Young, 1993).

In addition to subset pivotality and use of “Max” statistics, the additional assumption that there are no “logical constraints” among hypotheses (see Westfall and Tobias, 2007) is needed to assert that the algorithm is identical to closed testing using Max statistics. If there are logical constraints, power can be improved by restricting attention only to admissible subsets I, but in this case the computational shortcuts disappear, and one is back in the position of having to evaluate the tests for all intersections. The algorithm above can still be used, though, despite not being closed, as it provides a conservative procedure relative to the full closure.

3 Dispensing with Assumptions Altogether

While the subset pivotality condition is easily satisfied in many cases, including the general multivariate regression model with location-shift multivariate (possibly nonnormal) distributions (Westfall and Young, 1993, p. 123), researchers have questioned the assumption. In this section, we show how one can dispense with assumptions altogether, yet subset pivotality remains valid, and a powerful and flexible class of multiple testing procedures is obtained.

Let us clarify. By “dispensing with assumptions altogether,” we specifically mean “dispensing with all assumptions about the data-generating mechanism.” We will not assume the distributions lie in the location-shift family, or in any other family. We won't even assume that the data necessarily arise from a random process. This framework is similar to the simple falsification approach to non-multiplicity corrected hypothesis testing. For example, to test $H_0 : \mu = 0$ using a t-test, the i.i.d. $N(0; \sigma^2)$ assumptions all form the null hypothesis. No assumption is needed otherwise if you only want to control the type I error rate. Rejection of H_0 in this setting does not necessarily imply $\mu \neq 0$; rather, it implies rejection of independence, identical distributions, normality, or $\mu = 0$ (or, by the central limit theorem, in large samples it approximately implies rejection of independence, identical distributions, or $\mu = 0$). The conclusion $\mu \neq 0$ follows only if one is willing to accept all the other assumptions. The “other assumptions” are what we avoid altogether: everything will be embedded within the null hypotheses. We do this by using permutation tests, which allows for elegant theory, but a similar theory might be developed using other tests.

This approach of embedding all assumptions into the null hypothesis may not be ideal, from a research standpoint, because rejection of the null hypothesis may give the researcher little information as to why the hypothesis was rejected. Nevertheless, a major contribution of this paper is the development of this point of view in the multiple testing arena. We show, despite the very minimal setup, that a reasonably powerful and flexible class of multiple testing procedures is obtained.

We adopt the following framework for the data structure, hypotheses, and test statistics. The first element of the framework is that we have multivariate G-sample data. Such data abound in biometrical research, from adverse events data in clinical trials, to animal carcinogenicity data, to gene expression data. The second element is that the hypotheses of interest are that the treatments have no effect on the data, i.e., that the data are exchangeable, and the third concerns the form of the test statistics.

In fairness, the elements of the “framework” are restrictive; for example, if the researcher is only interested in location shift effects, then he or she is not interested in permutation tests. However, as stated above, the elements of the “framework” are somewhat different from the probabilistic assumptions that are usually made about data-generating processes. Specific elements of the framework are as follows:

1. The data are multivariate m-dimensional data vectors from G groups, $\boldsymbol{y}_{11}, \ldots, \boldsymbol{y}_{1n_1}, \ldots, \boldsymbol{y}_{G1}, \ldots, \boldsymbol{y}_{Gn_\mathrm{G}}$. Each $\boldsymbol{y}_{gj}$ vector is comprised of m elements, which may be character, numeric, or mixed, hence missing values are allowed. The data need not be generated by any random mechanism. For $I \subseteq \{1, \ldots, m\}$, let $\boldsymbol{y}^{\mathrm{I}}{}_{gj}$ denote the subvector of $\boldsymbol{y}_{gj}$ whose components are the elements of I.

2. The hypotheses of interest are

$$H_i\text{: the distribution of } \left\{\boldsymbol{Y}_{11}^{\{i\}}, \ldots, \boldsymbol{Y}_{1n_1}^{\{i\}}, \ldots, \boldsymbol{Y}_{G1}^{\{i\}}, \ldots, \boldsymbol{Y}_{Gn_G}^{\{i\}}\right\} \text{ is exchangeable.}$$

   Implicit in the hypothesis is an assumption of randomness; the hypothesis is equivalent to the statement that the observed data values

$$\boldsymbol{y}_{11}^{\{i\}}, \ldots, \boldsymbol{y}_{1n_1}^{\{i\}}, \ldots, \boldsymbol{y}_{G1}^{\{i\}}, \ldots, \boldsymbol{y}_{Gn_G}^{\{i\}}$$

   are realizations from an exchangeable random process (Pesarin, 2001, p. 5).

3. H_i is tested using a real-valued test statistic that is a function only of the data in variable i:

$$T_i = T_i\left\{\boldsymbol{Y}_{11}^{\{i\}}, \ldots, \boldsymbol{Y}_{1n_1}^{\{i\}}, \ldots, \boldsymbol{Y}_{G1}^{\{i\}}, \ldots, \boldsymbol{Y}_{Gn_G}^{\{i\}}\right\},$$

   Without loss of generality, larger values of T_i suggest non-exhangeability.

With these elements, closed testing can be performed to control the FWE, entirely free of probabilistic assumptions about the data-generating process, other than what is embedded in the null hypotheses. Permutation testing is used. Let $n = \sum n_g$, and suppose the data values

$$\boldsymbol{Y}^{\{i\}} := \left\{\boldsymbol{Y}_{11}^{\{i\}}, \ldots, \boldsymbol{Y}_{1n_1}^{\{i\}}, \ldots, \boldsymbol{Y}_{G1}^{\{i\}}, \ldots, \boldsymbol{Y}_{Gn_G}^{\{i\}}\right\}$$

are observed to be $\boldsymbol{y}^{\{i\}} := \left\{\boldsymbol{y}_1^{\{i\}}, \ldots, \boldsymbol{y}_n^{\{i\}}\right\}$, irrespective of specific group labels ($g = 1, \ldots, G$) and replicate labels ($j = 1, \ldots, n_g$). Specifically, letting $\mathcal{B}(y_1, \ldots, y_n)$ denote the permutation orbit of $\boldsymbol{y}$,

$$\mathcal{B}(y_1, \ldots, y_n) := \{x_1, \ldots, x_n | x_1, \ldots, x_n \text{ is a permutation of } y_1, \ldots, y_n\},$$

the conditioning set is $\boldsymbol{Y}^{\{i\}} \in \mathcal{B}(\boldsymbol{y}^{\{i\}})$.

Given H_i and $\boldsymbol{Y}^{\{i\}} \in \mathcal{B}(\boldsymbol{y}^{\{i\}})$, all $n!$ permutations of the data vector $\left\{\boldsymbol{y}_1^{\{i\}}, \ldots, \boldsymbol{y}_n^{\{i\}}\right\}$ are equally likely. Let $\boldsymbol{Y}_{11}^{*\{i\}}, \ldots, \boldsymbol{Y}_{1n_1}^{*\{i\}}, \ldots, \boldsymbol{Y}_{G1}^{*\{i\}}, \ldots, \boldsymbol{Y}_{Gn_G}^{*\{i\}}$ denote a randomly selected permutation among the $n!$, and define $t_\alpha^i\left(\boldsymbol{y}^{\{i\}}\right)$ by

$$t_\alpha^i\left(\boldsymbol{y}^{\{i\}}\right) = \min\left\{t : P\left(T_i\left\{\boldsymbol{Y}_{11}^{*\{i\}}, \ldots, \boldsymbol{Y}_{1n_1}^{*\{i\}}, \ldots, \boldsymbol{Y}_{G1}^{*\{i\}}, \ldots, \boldsymbol{Y}_{Gn_G}^{*\{i\}}\right\} \geq t\right) \leq \alpha\right.$$

if such a t exists, and $t_\alpha^i\left(\boldsymbol{y}^{\{i\}}\right) = \infty$ otherwise. The rejection rule is

$$\text{reject } H_i \text{ if } T_i \geq t_\alpha^i\left(\boldsymbol{y}^{\{i\}}\right).$$

Equivalently, since the permutation-based p-value is $p_i\left(\boldsymbol{y}^{\{i\}}\right) = P\left(T_i^* \geq t_i | H_i, \boldsymbol{Y}^{\{i\}} \in B\left(\boldsymbol{y}^{\{i\}}\right)\right)$ the rejection rule is

$$\text{reject } H_i \text{ if } p_i\left(\boldsymbol{Y}^{\{i\}}\right) \leq \alpha.$$

Conditional on $\boldsymbol{Y}^{\{i\}} \in \mathcal{B}(\boldsymbol{y}^{\{i\}})$, the rejection rule has Type I error rate $\leq \alpha$, with strict inequality likely due to the discreteness of the distribution of $T_i\left\{\boldsymbol{Y}_{11}^{*\{i\}}, \ldots, \boldsymbol{Y}_{1n_1}^{*\{i\}}, \ldots, \boldsymbol{Y}_{G1}^{*\{i\}}, \ldots, \boldsymbol{Y}_{Gn_G}^{*\{i\}}\right\}$. Since the Type I error rate is $\leq \alpha$ conditional on $\boldsymbol{Y}^{\{i\}} \in \mathcal{B}(\boldsymbol{y}^{\{i\}})$, it is also $\leq \alpha$ unconditionally.

To test all hypotheses using closure entails testing subsets $H_I = \cap_{i \in I} H_i$. Note that H_I implies marginal exchangeability of $\boldsymbol{Y}^{\{i\}}$, $i \in I$, but not of $\boldsymbol{Y}^I$. A clarifying example is as follows: let $\boldsymbol{Y}_{11}, \ldots, \boldsymbol{Y}_{1n}$ be i.i.d. bivariate normal with mean vector zero and identity covariance matrix, and let $\boldsymbol{Y}_{21}, \ldots, \boldsymbol{Y}_{2n}$ be i.i.d. bivariate normal with mean vector zero and covariance matrix

$$\begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}.$$

The component distributions are exchangeable but the joint distribution is not, and this is an example of a probability model in the set $H_{\{1,2\}}$.

Since the marginal models are specified but the joint models are not, the *p*-values $P(\max_{i\in I} T_i \geq \max_{i\in I} t_i \mid H_I)$ needed for testing the intersections are not easily determined. However, a Bonferroni-like inequality can be used: let the critical value for testing H_I at level α be

$$c^B_{I,\alpha}\left(\boldsymbol{y}^I\right)=\min\left\{c:\sum_{i\in I}P\left(T_i\geq c|H_i,\boldsymbol{Y}^{\{i\}}\in B\left(\boldsymbol{y}^{\{i\}}\right)\right)\leq\alpha\right\}$$

if such a c exists, define $c^B_{I,\alpha}\left(\boldsymbol{y}^I\right)=\infty$ otherwise. Then the rule $\max_{i\in I} T_i \geq c^B_{I,\alpha}$ provides an α-level test for H_I since

$$\begin{aligned}&P\left(\max_{i\in I}T_i\geq c^B_{I,\alpha}\left(\boldsymbol{y}^I\right)|H_I,\quad \boldsymbol{Y}^I\in B\left(\boldsymbol{y}^I\right)\right)\\&\leq\sum_{i\in I}P\left(T_i\geq c^B_{I,\alpha}\left(\boldsymbol{y}^I\right)|H_I,\quad \boldsymbol{Y}^I\in B\left(\boldsymbol{y}^I\right)\right)\\&=\sum_{i\in I}P\left(T_i\geq c^B_{I,\alpha}\left(\boldsymbol{y}^I\right)|H_i,\quad \boldsymbol{Y}^{\{i\}}\in B\left(\boldsymbol{y}^{\{i\}}\right)\right)\\&\leq\alpha.\end{aligned}$$

The *p*-value for this test is

$$p^B_I=\sum_{i\in I}P\left(T_i\geq\max_{i\in I}t_i|H_i,\boldsymbol{Y}^{\{i\}}\in B\left(\boldsymbol{y}^{\{i\}}\right)\right);$$

note that $p^B_I\leq\alpha$ if and only if $\max_{i\in I} t_i\geq c^B_{I,\alpha}\left(\boldsymbol{y}^I\right)$.

FWE-controlling tests for the H_i follow closure and the “Main algorithm” of Section 2.3, with identical shortcuts resulting from use of “Max” tests and the resulting monotonicity of *p*-values. As in Section 2.3, suppose the observed test statisticsare $t_1 \geq \cdots \geq t_m$, corresponding to hypotheses $H_1, \ldots, H_m$ (again ordered in this way without loss of generality). The main algorithm becomes, in this case,

1. Reject H_1 if

$$p^B_{\{1,\ldots,m\}}\leq\alpha.$$

2. Reject H_2 if

$$p^B_{\{1,\ldots,m\}}\leq\alpha$$

and if

$$p^B_{\{2,\ldots,m\}}\leq\alpha$$

j.: Continuing in this fashion, the rule is reject H_j if

$$p^B_{\{1,\ldots,m\}}\leq\alpha$$

and

$$p^B_{\{2,\dots,m\}} \le \alpha$$

and … and

$$p^B_{\{j,\dots,m\}} \le \alpha.$$

As before, the adjusted *p*-values are $\tilde{p}_j=\max_{i\le j} p^B_{\{i,\dots,m\}}$.

This method is the "discrete Bonferroni method" described by Westfall and Wolfinger (1997), which has been hard-coded in PROC MULTTEST of SAS/STAT since 1996. What is unique about the above presentation is the generalization to arbitrary test statistics (Westfall and Wolfinger's results arise when the $-t_i$ is a marginal *p*-value, either exact or approximate). Use of *p*-values results in balance, where no particular hypotheses are favored. However, the more general framework allows deliberate weighting to favor certain hypotheses *a priori*. For example, if the supports of the permutation distributions of the T_i are completely disjoint, the algorithm reduces to the "*a priori ordered*" testing procedure described in, e.g., Maurer et al. (1995) and Kropf et al. (2004).

What is also unique about the above presentation is the assertion that absolutely no assumptions are needed concerning the data-generating process, not even independence, and certainly not subset pivotality. Curiously, even though subset pivotality is considered "restrictive," subset pivotality holds in this general case where no assumptions are made: the joint distribution of the test statistics $\{T_i : i \in I\}$ is the same under both H_I and H_S, following specifically from element 3 of the framework described above in this section. Thus, subset pivotality is hardly restrictive, since no assumptions, other than what is embedded in the null hypotheses, are made. The only problem is that H_I is not easily characterized, so that *p*-values based on the joint distributions cannot be easily calculated without further assumptions. Hence the assumption that is questionable is not subset pivotality, but the assumption needed to calculate *p*-values under H_I. That assumption will be given in Section 4. For now, we use the conservative Bonferroni-based approximation to the *p*-values based on the joint distributions, as shown above.

Example Consider the adverse event data set provided by Westfall et al. (1999, p. 243). There are $G = 2$ groups, control and treatment, with $n_g = 80$ patients in each group, and $m = 28$ adverse event indicator variables per patient. Null hypotheses are that the adverse events indicator data are exchangeable in the combined treatment control sample, tested using Fisher exact upper tailed *p*-values, with smaller *p*-values indicating more adverse events in the treatment group. Unadjusted and adjusted *p*-values for the five most significant adverse events (AEs) are as follows:

Especially noteworthy is the adjusted *p*-value for the most extreme adverse event: since $0.0025 \approx 3 \times 0.0008$, the effective number of tests is 3, not 28. The discrete Bonferroni method is thus vastly superior to the ordinary Bonferroni method, for which the adjusted *p*-value would have been $28 \times 0.0008 = 0.0224$. Again, this benefit comes at no expense of extra assumptions, since no assumptions are made concerning the data generating process.

Westfall and Soper (2001) note that this method automatically adjusts for selection effects concerning the observed variables. Here, the observed adverse events are self-reported, thus there is concern about a possible selection effect concerning the particular 28 adverse events that were reported. This is not an issue though when one realizes that in the collection of possible adverse events that could be reported, the total reports are 0 for all but the 28 in this study. Specifically, suppose there are 100 possible reportable AEs, 72 of which produce no reports. The permutation distributions for those 72 events with no reports place 100% of their mass on the *p*-value 1.0, hence the discrete Bonferroni analysis of all 100 AEs is identical to the analysis of the 28 where reports are received. However, the ordinary Bonferroni correction would change from 28 × 0.0008 = 0.0224 to 100 × 0.0008 = 0.080.

4 Improving Power Using Joint Distributions

The method described in Section 3 does not utilize joint distribution information, and therefore can be improved. To utilize this information, a joint exchangeability assumption is needed, as noted by Westfall (2003), and Calian et al. (2008). Consider the same setup as the previous section, with one assumption, rather than none, about the data-generating process. This is the assumption that people often confuse, erroneously, with subset pivotality:

Assumption C If for $I, J \subseteq \{1, \ldots, m\}$ the distribution of $\left\{\boldsymbol{Y}^{I}_{11}, \ldots, \boldsymbol{Y}^{I}_{1n_1}, \ldots, \boldsymbol{Y}^{I}_{G1}, \ldots, \boldsymbol{Y}^{I}_{Gn_G}\right\}$ is exchangeable in its n components, and the distribution of $\left\{\boldsymbol{Y}^{J}_{11}, \ldots, \boldsymbol{Y}^{J}_{1n_1}, \ldots, \boldsymbol{Y}^{J}_{G1}, \ldots, \boldsymbol{Y}^{J}_{Gn_G}\right\}$ is exchangeable in its n components, then the distribution of $\left\{\boldsymbol{Y}^{I\cup J}_{11}, \ldots, \boldsymbol{Y}^{I\cup J}_{1n_1}, \ldots, \boldsymbol{Y}^{I\cup J}_{G1}, \ldots, \boldsymbol{Y}^{I\cup J}_{Gn_G}\right\}$ is exchangeable in its n components.

In particular, the assumption implies that $\cap_{i\in I} H_i = H_I$: {the distribution of $\left\{\boldsymbol{Y}^{I}_{11}, \ldots, \boldsymbol{Y}^{I}_{1n_1}, \ldots, \boldsymbol{Y}^{I}_{G1}, \ldots, \boldsymbol{Y}^{I}_{Gn_G}\right\}$ is exchangeable}. Like all assumptions, this one is questionable; a simple counter-example with two-group bivariate normal data is given in Section 3. However, several points can be made concerning the palatability of the assumption. First, the class of allowable models satisfying Assumption C is substantially more general than the multivariate location-shift class of models, allowing non-normality; while the multivariate location-shift class of models is very commonly used, often assuming normality. Second, it is perhaps not unrealistic to assume that if there is no treatment effect for each of variables 1, 2 individually, then the joint distribution of $\left\{\boldsymbol{Y}^{\{1,2\}}_{11}, \ldots, \boldsymbol{Y}^{\{1,2\}}_{1n_1}, \ldots, \boldsymbol{Y}^{\{1,2\}}_{G1}, \ldots, \boldsymbol{Y}^{\{1,2\}}_{Gn_G}\right\}$ is exchangeable. Third, even if this is not a realistic assumption, failure of the assumption might imply excess Type I errors. But if the treatment really does affect the response, then the researcher might take comfort in conclusions of statistical significance, while acknowledging that due to assumption failure, the effect might be on correlation structure rather than on specific variables. Fourth, we reiterate that no assumption of independence is needed.

The benefit of Assumption C is that the exact *p*-value for H_I can be calculated: the conditional *p*-value

$$p_I := P\left(\max_{i\in I} T_i \geq \max_{i\in I} t_i | H_I, \boldsymbol{Y}^I \in B\left(\boldsymbol{y}^I\right)\right)$$

is free of the H_I-distribution of the data; p_I is specifically equal to the proportion of the n! permutations yielding

$$\max_{i\in I} T_i\left\{\boldsymbol{Y}^{*\{i\}}_{11},\ldots,\boldsymbol{Y}^{*\{i\}}_{1n_1},\ldots,\boldsymbol{Y}^{*\{i\}}_{G1},\ldots,\boldsymbol{Y}^{*\{i\}}_{Gn_G}\right\}\geq \max_{i\in I} t_i.$$

Note also that $p_I \leq p_I^B$, where p_I^B is defined in the previous section, hence incorporating dependence information can provide greater power.

See Puri and Sen (1971) and Pesarin (2001) for further details on multivariate permutation tests.

As in Section 3, subset pivotality holds, and use of the “Max” statistic imply that the main algorithm can be used directly. In this case it becomes

1. Reject H_1 if

$$p_{\{1,\ldots,m\}} \leq \alpha.$$

2. Reject H_2 if

$$p_{\{1,\ldots,m\}} \leq \alpha$$

and if

$$p_{\{2,\ldots,m\}} \leq \alpha$$

j.: Continuing in this fashion, the rule is reject H_j if

$$p_{\{1,\ldots,m\}} \leq \alpha$$

and

$$p_{\{2,\ldots,m\}} \leq \alpha$$

and … and

$$p_{\{j,\ldots,m\}} \leq \alpha.$$

As before, the adjusted *p*-values are $\tilde{p}_j = \max_{i\leq j} p_{\{i,\ldots,m\}}$.

5 Concluding Remarks

Remark 1 Often, complete enumeration of the *n*! permutations is infeasible, so the *p*-values are instead approximated by randomly sampling permutations:

1. Generate a resampled data set $\boldsymbol{Y}^*_{gj}, \quad g=1,\ldots,G,\ j=1,\ldots,n_i$, a *without replacement* sample from the observed vectors $\{\boldsymbol{y}_{11},\ldots,\boldsymbol{y}_{1n_1},\ldots,\boldsymbol{y}_{G1},\ldots,\boldsymbol{y}_{Gn_G}\}$.
2. Compute the statistics T_i^* from the $\boldsymbol{Y}^*_{gj}$.
3. Check whether

$$\max_{i \in I} \quad T_i^* \geq \max_{i \in I} t_i.$$

4. Repeat 1.–3. a large number B (in the millions preferably) of times. The exact permutation p-value p_I is approximately (within binomial simulation error) the proportion of the B samples where $\max_{i \in I} \quad, T_i^* \geq \max_{i \in I} t_i$.

This method has been hard-coded in PROC MULTTEST of SAS/STAT since the inception of the PROC in 1992.

Consider the example in Section 3. The adjusted p-values using the joint distributions are shown below, all of which are calculated using 5,000,000 draws from the multivariate permutation distribution, along with the discrete Bonferroni adjustments from Section 3 for comparison.

The specific benefit of making Assumption C is the ability to incorporate correlation information; this benefit is shown by the smaller adjusted p-values. However, the main benefit shown by the discrete Bonferroni method comes at the expense of no additional assumptions whatsoever, and we reiterate that subset pivotality holds in that case, as described in Section 3.

Remark 2 The “joint distribution” method is exact, in the sense that all composite hypotheses in the closure are tested using exact, distribution-free permutation tests. The method also incorporates all correlation information between variables. It is perhaps surprising that exact, correlation-incorporating tests are possible, even when the multivariate dimension m is much greater than the sample size n.

Remark 3 Assumption C is equivalent to that assumed by Korn et al. (2004) for control of the False Discovery Proportion. Although the Korn et al. procedure is necessarily more computationally challenging when allowing for some false discoveries, it has a similar structure as the algorithm in Section 4, and it reduces to that algorithm when allowing no false discoveries.

To conclude, subset pivotality is an assumption made strictly for the computational convenience. It need not be restrictive: as shown in Section 3, subset pivotality holds in the most minimal setup. The assumption that people seem to object to is not subset pivotality, but the assumption that marginal exchangeability implies joint exchangeability, assumption C. However, we noted (a) assumption C might not be objectionable, and (b) as the analysis of the adverse events data show, even assumption C is not crucial. If objectionable, one can revert to the discrete Bonferroni method, which makes no assumptions. In cases with moderate dependence structure and highly discrete permutation distributions, the discrete Bonferroni method can provide a more important benefit than the benefit obtained by incorporating of joint distribution information, for which the potentially objectionable assumption is needed.

Acknowledgments

This research was supported in part by the Intramural Research Program of the National Institutes of Health and the National Institute of Child Health and Human Development.

References

Calian V, Li D, Hsu JC. Partitioning to uncover conditions for permutation tests to control multiple testing error rates. Biometrical Journal. 2008; 50:756–766. this issue. [PubMed: 18932135]

Dmitrienko A, Offen W, Westfall PH. Gatekeeping strategies for clinical trials that do not require all primary effects to be significant. Statistics in Medicine. 2003; 22:2387–2400. [PubMed: 12872297]

Holm S. A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics. 1979; 6:65–70.

Korn EL, Troendle JF, McShane LM, Simon R. Controlling the number of false discoveries: application to high-dimensional genomic data. Journal of Statistical Planning and Inference. 2004; 124:379–398.

Kropf S, Läuter J, Eszlinger M, Krohn K, Paschke R. Nonparametric multiple test procedures with data-driven order of hypotheses and with weighted hypotheses. Journal of Statistical Planning and Inference. 2003; 125:31–47.

Marcus R, Peritz E, Gabriel KR. On closed testing procedure with special reference to ordered analysis of variance. Biometrika. 1976; 63:655–660.

Maurer, W.; Hothorn, LA.; Lehmacher, W. Multiple comparisons in drug clinical trials and preclinical assays: A-priori ordered hypotheses. In: Vollmar, J., editor. Biometrie in der chem.-pharm. Industrie. Vol. 6. Fischer-Verlag; Stuttgart: 1995. p. 2-18.

Pesarin, F. Multivariate Permutation Tests: With Applications in Biostatistics. Wiley; Chichester: 2001.

Puri, ML.; Sen, PK. Nonparametric Methods in Multivariate Analysis. Wiley; New York: 1971.

Romano JP, Shaikh AM, Wolf M. Formalized data snooping based on generalized error rates. Econometric Theory. 2008; 24:404–447.

Westfall, PH.; Tobias, R.; Rom, D.; Wolfinger, R.; Hochberg, Y. Multiple Comparisons and Multiple Tests using SAS®. SAS Institute Inc.; Cary, NC: 1999.

Westfall, PH.; Young, SS. Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment. Wiley; New York: 1993.

Westfall PH, Wolfinger RD. Multiple tests with discrete distributions. The American Statistician. 1997; 51:3–8.

Westfall PH, Soper KA. Using priors to improve multiple animal carcinogenicity tests. Journal of the American Statistical Association. 2001; 96:827–834.

Westfall PH, Tobias RD. Multiple testing of general contrasts: truncated closure and the extended Shaffer-Royen method. Journal of the American Statistical Association. 2007; 102:487–494.

Westfall PH. Comment on “Resampling-based Multiple Testing for Microarray Data Analysis,” Y. Ge, S. Dudoit and T. P. Speed. Test. 2003; 12:60–65.

Table 1

Unadjusted and discrete Bonferroni-adjusted *p*-values for adverse event data.

***p*-value**	**AE1**	**AE8**	**AE6**	**AE5**	**AE10**
Unadjusted	0.0008	0.0293	0.0601	0.2213	0.2484
Discrete Bonferroni adjusted	0.0025	0.1587	0.3321	1.0000	1.0000

Table 2

Table 1, with joint distribution adjusted *p*-values.

***p*-value**	**AE1**	**AE8**	**AE6**	**AE5**	**AE10**
Unadjusted	0.0008	0.0293	0.0601	0.2213	0.2484
Discrete Bonferroni adjusted	0.0025	0.1587	0.3321	1.0000	1.0000
Joint distribution adjusted	0.0021	0.1335	0.2605	0.6278	0.9275