Published in final edited form as:

J Stat Plan Inference. 2011 June 1; 141(6): 2021–2029. doi:10.1016/j.jspi.2010.12.012.

Permutational Multiple Testing Adjustments With Multivariate Multiple Group Data

James F. Troendle^{*,†,‡} and Peter H. Westfall[§]

* Biostatistics and Bioinformatics Branch of the Division of Epidemiology, Statistics, and Prevention Research of the Eunice Kennedy Shriver National Institute of Child Health and Human Development, NIH/DHHS

§ Texas Tech University, Lubbock TX

Abstract

We consider the multiple comparison problem where multiple outcomes are each compared among several different collections of groups in a multiple group setting. In this case there are several different types of hypotheses, with each specifying equality of the distributions of a single outcome over a different collection of groups. Each type of hypothesis requires a different permutational approach. We show that under a certain multivariate condition it is possible to use closure over all hypotheses, although intersection hypotheses are tested using Boole's inequality in conjunction with permutation distributions in some cases. Shortcut tests are then found so that the resulting testing procedure is easily performed. The error rate and power of the new method is compared to existing competitors through simulation of correlated data. An example is analyzed, consisting of multiple adverse events in a clinical trial.

Keywords

Adverse Events; Closed Testing; Exchangeability; Familywise Error Rate; Power

1 Introduction

It is common practice in analysis of biomedical data to be interested in testing a set of hypotheses, raising the issue of multiple comparison adjustment. Correlation and discreteness make it important to use resampling in the adjustments for multiplicity (Westfall and Young, 1993). An example considered here is the case of multiple outcomes, where each outcome is of interest in itself. This basic problem has been studied, with several good resampling solutions available (Troendle, 1995; Troendle, 1996; Westfall and Troendle, 2008). The problem is more complicated if one has multiple hypotheses of interest for each outcome. Since in the ideal case of two samples with a null hypothesis of exchangeability the permutation test is exact, permutation adjustments are preferred over bootstrap adjustment when applicable. However, more than two groups can lead to different types of hypotheses, which require different permutational approaches to adjustment.

[†]Currently at Office of Biostatistics Research of the Division of Cardiovascular Sciences of the National Heart, Lung, and Blood Institute, NIH/DHHS, Bld RKL2 Room 9195, Bethesda, MD 20892; jt3t@nih.gov.

[‡]This research was supported in part by the Intramural Research Program of the NIH, NICHD.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Permutation tests are popular for biostatistical applications quite generally, because they are exact, guaranteeing type I error control for non-normal distributions regardless of sample size. Desirable properties of permutation tests have lead some to advocate their use for multiple comparisons in the multiple group ANOVA, using the global permutation distribution over all groups to obtain multiplicity adjustments for the pairwise comparisons. Unfortunately, the exactness of permutation univariate tests does not imply exactness of the multiple comparison procedure when they are conducted using global randomizations, as shown by Petrondas and Gabriel (1983).

In this paper we show how to perform permutation tests for multiple comparisons and multiple variables simultaneously, while avoiding the problems associated with the use of global permutation distributions. The intended application is clinical trials where there are many variables and few groups (say, five or less); otherwise some of the methods become computationally prohibitive. We use separate permutation distributions for each type of hypothesis under consideration, identifying stepwise and single-step procedures that guarantee control of the probability of any type I error, or familywise error rate (FWE). All procedures are derived from the closure principal of Marcus, Peritz, and Gabriel (1976), and control of the FWE follows automatically. Further, the closed testing scheme is shown to be computable, as shortcuts following from monotonicity properties allow that the testing is carried out in step-down fashion, i.e. follows the order of the observed test statistics. Finally, the procedures account for joint correlation and distributional characteristics. An example with binary data shows that the method offers substantial gains in power over standard parametric closed testing methods.

The paper proceeds as follows. Section 2 considers the case where two classes of hypotheses require different types of permutation of the data. Section 3 gives an example of adverse events in a clinical trial. Section 4 presents simulations comparing the error rates and power under various null and alternative configurations. Finally, section 5 contains recommendations.

2 Null Hypotheses That Imply Different Models

2.1 Closed Testing

We consider the problem of testing null hypotheses, $H_1, \dots, H_k$, while controlling the probability of any type I error over the k decisions to be $\leq \alpha$ for some prespecified $0 < \alpha < 1$. In order to obtain multiple comparison procedures that control the FWE, the closure principle of Marcus, Peritz, and Gabriel (1976) is often applied. A closed testing procedure (CTP) requires that all intersection hypotheses $H_I = \bigcap_{i \in I} H_i$, for $I \subseteq S := \{1, \dots, k\}$ be tested by base tests. Then a CTP is a multiple comparison procedure that rejects an arbitrary intersection hypothesis H iff H_- is rejected for every intersection hypothesis $H_- \subseteq H$. If the base tests of H_I are level α , then the CTP controls the FWE at level α .

If $H_I \neq H_J$ for all $I, J \in S$ with $I \neq J$, there are $2^k - 1$ unique intersection hypotheses H_I , and the hypotheses H_i are said to obey the “free combinations” condition. In the “restricted combination” case, $H_I = H_J$ for some $I, J \in S$ with $I \neq J$, and there are fewer than $2^k - 1$ unique intersection hypotheses H_I . In the free combinations case there is a simplifying shortcut that allows testing to proceed using the k ordered test statistics rather than the $2^k - 1$ subsets; this will be shown below. In the restricted combinations case, such shortcuts do not occur generally. Nevertheless, a procedure that treats the hypotheses as free is necessarily conservative relative to the full closure, since if a single hypothesis $H_I (= H_J$ for some $I \neq J)$ is tested twice, using test statistics $T_I \neq T_J$, the frequency of analyses where the test is rejected twice is uniformly as small or smaller than the frequency of analyses where H_I is tested once and rejected using T_I . Hence, the FWE is also controlled when the hypotheses

are treated as free, but more power is available by exploiting logical restrictions when they exist.

In either case, we will treat the hypotheses as free and construct multiple testing procedures based on the CTP, where the base tests of H_I are permutation tests; the resulting procedures are more robust than parametric alternatives, and FWE control is guaranteed by virtue of closure and the exactness of permutation tests. Further, implicit incorporation of characteristics such as discreteness and correlation structure can make the procedures more powerful than standard parametric methods.

2.2 Setup

Assume that we have observations on $\mathbf{Y} = \{\mathbf{Y}_{11}, \dots, \mathbf{Y}_{1n_1}, \mathbf{Y}_{21}, \dots, \mathbf{Y}_{2n_2}, \dots, \mathbf{Y}_{G1}, \dots, \mathbf{Y}_{Gn_G}\}$, $n = \sum n_g$ independent multivariate q dimensional random vectors from G groups. The vectors may contain character and/or missing values. For $I \subseteq Q = \{1, \dots, q\}$, let $\mathbf{Y}_{gj}^I$ denote the subvector of $\mathbf{Y}_{gj}$ whose component positions are indicated by the elements of I , with analogous definition for $\mathbf{Y}^I$ as submatrix of $\mathbf{Y}$. Let $F_g^{(i)}$ be the distribution of component i from group g . Hypotheses of interest are equality of group distributions (two or more), for each component. Let c be the number of hypotheses of interest on each component. For a simple example with $c = 2$, the component hypotheses might represent a comparison between control and high dose, and a comparison between control and low dose.

For $l = 1, \dots, c$, let $A_l \subseteq \{1, \dots, G\}$ and let $\mathbf{Y}_l^I = \{\mathbf{Y}_{gj}^I; g \in A_l, j=1, \dots, n_g\}$ denote the data in groups A_l . The hypotheses to be tested specify equality of distributions of component i in groups A_l :

$$H_{l,i}: F_g^{(i)} = G_l^{(i)} \text{ for } g \in A_l,$$

where $G_l^{(i)}$ denotes the common distribution. Note that $H_{l,i}$ implies component i data in groups A_l are exchangeable. Such hypotheses are also considered in Korn et al. (2004) and Westfall and Troendle (2008). Our example in Section 3 will concern testing of two two-sample hypotheses along with one hypothesis that implies that four groups are exchangeable.

Each $H_{l,i}$ is tested by a real valued test statistic that is a function only of the data in component i and groups A_l :

$$T_{l,i} = T_{l,i}(\mathbf{Y}_l^{(i)}).$$

Without loss of generality, assume larger values of $T_{l,i}$ suggest non-exchangeability.

Let $\mathbf{y}_l^{(i)}$ be the observed data on $\mathbf{Y}_l^{(i)}$, and define a permutation orbit

$$\mathcal{B}(\mathbf{y}_l^{(i)}) = \{\mathbf{x}_l^{(i)} | \mathbf{x}_l^{(i)} \text{ is a permutation of } \mathbf{y}_l^{(i)}\},$$

allowing possible duplicates so that there are $N_l!$ elements of $\mathcal{B}(\mathbf{y}_l^{(i)})$, where $N_l = \sum_{i \in A_l} n_i$. The conditioning event for hypothesis $H_{l,i}$ is $\mathcal{B}(\mathbf{y}_l^{(i)})$, defined as the event

$$B(\mathbf{y}_l^{(i)}) = \{\mathbf{Y}_l^{(i)} \in \mathcal{B}(\mathbf{y}_l^{(i)})\}.$$

Given $H_{l,i}$ and $B(\mathbf{y}_l^{(i)})$, all elements of $\mathcal{B}(\mathbf{y}_l^{(i)})$ are equally likely outcomes for $\mathbf{Y}_l^{(i)}$, see e.g., Strasser and Weber (1999). Hence, letting $P_{l,i}(\cdot)$ and $E_{l,i}(\cdot)$ denote probability and expectation calculated under $H_{l,i}$, we have

$$P_{l,i}(T_{l,i} \geq t | B(\mathbf{y}_l^{(i)})) = \left| \left\{ \mathbf{x}_l^{(i)} \in \mathcal{B}(\mathbf{y}_l^{(i)}) : T_{l,i}(\mathbf{x}_l^{(i)}) \geq t \right\} \right| / N_l!$$

Define

$$t_{l,i}^\alpha = t_{l,i}^\alpha(\mathbf{y}_l^{(i)}) = \begin{cases} \min \{t : P_{l,i}(T_{l,i} \geq t | B(\mathbf{y}_l^{(i)})) \leq \alpha\} & \text{if such a } t \text{ exists} \\ \infty & \text{otherwise} \end{cases}.$$

It follows that $h_t(\mathbf{y}_l^{(i)}) \equiv P_{l,i}(T_{l,i} \geq t_{l,i}^\alpha | B(\mathbf{y}_l^{(i)})) \leq \alpha$; hence the rejection rule

$$\text{reject } H_{l,i} \text{ if } T_{l,i} \geq t_{l,i}^\alpha$$

provides a test with level no greater than α conditional on the event $B(\mathbf{y}_l^{(i)})$. It also follows that $h_t(\mathbf{Y}_l^{(i)}) \leq \alpha$ with probability one. The test has level no greater than α unconditionally as well since

$$P_{l,i}(T_{l,i} \geq t_{l,i}^\alpha) = E_{l,i}[h_t(\mathbf{Y}_l^{(i)})] \leq \alpha.$$

2.3 Stepwise Multiple Comparison Procedure Based on Testing Intersection Hypotheses by Combining Permutation Tests Using Boole's Inequality

To extend the test given above in Section 2.2 to a multiple comparison procedure using closure, we need to test arbitrary intersection hypotheses of the form $H_I = \bigcap_{l \in L} \bigcap_{i \in I_l} H_{l,i}$. In the next section we will make an assumption about multivariate distributions in order to exploit the correlation fully, but for now we use Boole's inequality to obtain a conservative critical value for testing H_I at level α from the individual permutation tests of $H_{l,i}$.

Define the critical value for testing H_I to be

$$c_I^\alpha = c_I^\alpha(\mathbf{y}) = \min \left\{ c : \sum_{l \in L} \sum_{i \in I_l} P_{l,i}(T_{l,i} \geq c | B(\mathbf{y}_l^{(i)})) \leq \alpha \right\}, \quad (1)$$

if such a c exists, and define $c_{I'}^\alpha(\mathbf{y}) = \infty$ otherwise. Note that if $I'_1 \subseteq I'_2$ then $c_{I'_1}^\alpha \leq c_{I'_2}^\alpha$. This can be seen by considering the effect of enlarging the set I'_1 , which adds terms to the sum in (1). The effect of more terms in the sum in (1) is that the set in (1) is smaller, leading to a larger minimum. Define $h_c(\mathbf{y}_l^{(i)}) \equiv P_{H_{I'}}(T_{l,i} \geq c_{I'}^\alpha | B(\mathbf{y}_l^{(i)}))$.

Note that, given $B(\mathbf{y}_l^{(i)})$, the distribution of $T_{l,i}$ is determined by the $N_l!$ elements of $\mathcal{B}(\mathbf{y}_l^{(i)})$ whenever $H_{l,i}$ is true. But $H_{l,i}$ is true in particular when $H_{I'}$ is true. Hence, given $B(\mathbf{y}_l^{(i)})$, the distribution of $T_{l,i}$ is the same under $H_{l,i}$ as under $H_{I'}$. This property is a special case of the “subset pivotality condition” of Westfall and Young (1993, p. 42), and implies that

$$\sum_{l \in L} \sum_{i \in I_l} h_c(\mathbf{y}_l^{(i)}) \leq \alpha. \quad (2)$$

It follows that $\sum_{l \in L} \sum_{i \in I_l} h_c(\mathbf{Y}_l^{(i)}) \leq \alpha$ with probability one. The test which rejects $H_{I'}$ if

$$\max_{l \in L} \max_{i \in I_l} T_{l,i} \geq c_{I'}^\alpha$$

has level α since

$$\begin{aligned} P_{H_{I'}} \left(\max_{l \in L} \max_{i \in I_l} T_{l,i} \geq c_{I'}^\alpha \right) &\leq \sum_{l \in L} \sum_{i \in I_l} P_{H_{I'}}(T_{l,i} \geq c_{I'}^\alpha) \quad \text{by Boole's inequality} \\ &= E_{I'} \left[\sum_{l \in L} \sum_{i \in I_l} h_c(\mathbf{Y}_l^{(i)}) \right] \\ &\leq \alpha. \end{aligned}$$

A multiple comparison procedure can then be formed by following the principle of closure described in Section 2.1. For simplicity of notation, suppose the observed test statistics are $t_1 \geq t_2 \geq \dots \geq t_k$, corresponding to hypotheses $H_1, H_2, \dots, H_k$. To test H_1 , reject H_1 if

$$\max_{i \in I'} t_i \geq c_{I'}^\alpha \quad \text{for all } I' \text{ such that } \{1\} \subseteq I'.$$

However, if $\{1\} \subseteq I'$, then $\max_{i \in I'} t_i = t_1$, and the rejection region becomes

$$t_1 \geq \max_{I': \{1\} \subseteq I'} c_{I'}^\alpha = c_S^\alpha,$$

where the equality follows because of the monotone property of the critical constants defined in (1). To test H_2 , reject H_2 if

$$\max_{i \in I'} t_i \geq c_{I'}^\alpha \quad \text{for all } I' \text{ such that } \{2\} \subseteq I'.$$

For any such I' if $\{1\} \subseteq I'$, then $\max_{i \in I'} t_i = t_1$, and otherwise $\max_{i \in I'} t_i = t_2$. Partition the set $\{I': \{2\} \subseteq I'\}$ into $S_1 = \{I': \{1, 2\} \subseteq I'\}$ and $S_2 = \{I': \{2\} \subseteq I', \{1\} \subseteq S/I'\}$. Then the rejection region becomes

$$t_1 \geq c_{I'}^\alpha \quad \text{for all } I' \in S_1 \quad \text{and} \quad t_2 \geq c_{I'}^\alpha \quad \text{for all } I' \in S_2.$$

Again by monotonicity of $c_{I'}^\alpha$, the rejection region becomes

$$t_1 \geq c_S^\alpha \quad \text{and} \quad t_2 \geq c_{\{2, \dots, k\}}^\alpha.$$

Continuing in this fashion one obtains that to reject H_j , one requires

$$t_1 \geq c_S^\alpha \quad \text{and} \quad t_2 \geq c_{\{2, \dots, k\}}^\alpha \quad \text{and} \quad \dots \quad \text{and} \quad t_j \geq c_{\{j, \dots, k\}}^\alpha.$$

The final procedure can then be described compactly in the following algorithm based on the ordered test statistics and hypotheses.

Stepwise Permutational Boole's Algorithm

- 0 Set $j = 1$.
- 1 If $t_j \geq c_{\{j, \dots, k\}}^\alpha$ then reject H_j and continue with step 2, otherwise stop testing.
- 2 Increment $j \rightarrow j + 1$ and return to step 1 if $j \leq k$, otherwise stop testing.

Note that this procedure does not require any assumptions other than the basic framework described in Section 2.2. This means that the procedure can be applied to any multiple outcome problem with null hypotheses that imply exchangeability on a single outcome and use test statistics that are functions only of the relevant outcome in the relevant groups. Further, the method automatically adjusts for discreteness which can lead to some outcomes having little or no contribution to the multiplicity adjustment. As an example, consider the classic two independent sample case with Bernoulli outcomes. If one outcome had no observed successes for any individual in either group, then the p-value for that outcome would be 1.0 for every permutation, and the outcome would effectively drop out of the problem because the corresponding term in the sum in (1) would be zero at the minimizing value of c . In contrast, the Bonferroni procedure would still count that outcome in its adjustment. If there were several such outcomes, they would all effectively drop out of the multiplicity adjustment. Thus the method described here would be much more powerful than the Bonferroni procedure.

In applications where one would not want to favor any hypothesis over any other, one could calculate a raw p-value, P_i , for each individual hypothesis and set $T_i = -P_i$ for a more balanced application. On the other hand, the method described above applies to any test statistics, and can be devised with weights to allow more power to certain hypotheses. We

use the balanced treatment which sets $T_i = -P_i$ in an application to adverse events given in Section 3.

2.4 Stepwise Multiple Comparison Procedure Based on Testing Intersection Hypotheses by the Combination of Boole's Inequality and Multivariate Permutation Tests

The procedure described in Section 2.3 does not incorporate the correlation structure of the test statistics over the different components, and therefore can be improved. To improve that procedure, a joint distributional assumption is required. Denote by F_g^I the multivariate distribution of components I of group g .

Joint Distributional Assumption—If for $I, J \subseteq \{1, \dots, q\}$ and $A_I \subseteq \{1, \dots, G\}$ the distributions $\{F_g^I: g \in A_I\}$ are equal and the distributions $\{F_g^J: g \in A_I\}$ are equal, then the distributions $\{F_g^{I \cup J}: g \in A_I\}$ are equal.

An example where the assumption does not hold, also given in Westfall and Troendle (2008), is for two sample multivariate normal data with equal variance but unequal correlation in the groups. In that case the component distributions are equal without the joint distribution being so. The first part of the Joint Distributional Assumption is therefore violated. It is worth noting that the assumption does not imply independence between components. The assumption holds in the MANOVA model with i.i.d. error vectors; see e.g. Westfall and Young (1993, p. 122–3).

The joint distribution assumption (JDA) implies exchangeability of the I -dimensional column data vectors Y_{gj}^I for $g \in A_I$ under the hypothesis $\cap_{i \in I} H_{l,i}$. Under the JDA, one can use multivariate permutations of the subvectors, while tying together different test types by Boole's inequality, to obtain tests of intersection hypotheses in a CTP. To do this define $B(\mathbf{y}_I^I)$ as the event:

$$B(\mathbf{y}_I^I) = \{Y_I^I \in \mathcal{B}(\mathbf{y}_I^I)\}.$$

Then the critical value for testing H_I is defined as

$$d_{I'}^\alpha = d_{I'}^\alpha(\mathbf{y}) = \min \left\{ d: \sum_{l \in L} P_{H_{l,I_l}} \left(\max_{i \in I_l} T_{l,i} \geq d | B(\mathbf{y}_I^I) \right) \leq \alpha \right\}, \quad (3)$$

if such a d exists, $d_{I'}^\alpha(\mathbf{y}) = \infty$ otherwise, and where $H_{l,I_l} = \cap_{i \in I_l} H_{l,i}$. Here we have that, analogous to the monotonicity condition for $c_{I'}^\alpha$, found in Section 2.3, if $I'_1 \subseteq I'_2$ then $d_{I'_1}^\alpha \leq d_{I'_2}^\alpha$. Define $h_d(\mathbf{y}_I^I) \equiv P_{H_{I'}} \left(\max_{i \in I_l} T_{l,i} \geq d_{I'}^\alpha | B(\mathbf{y}_I^I) \right)$.

Note that, given $B(\mathbf{y}_I^I)$, the distribution of $\{T_{l,i}: i \in I_l\}$ is determined by the $N_l!$ elements of $\mathcal{B}(\mathbf{y}_I^I)$ whenever H_{l,I_l} is true. But, as in the univariate case, H_{l,I_l} is true when H_I is true. Hence the distribution of $\{T_{l,i}: i \in I_l\}$ is the same under H_{l,I_l} as under H_I ; in either case it is determined by the $N_l!$ elements of $\mathcal{B}(\mathbf{y}_I^I)$. The property implies that

$$\sum_{l \in L} h_d(\mathbf{Y}_l^{I_l}) \leq \alpha. \quad (4)$$

It follows that $\sum_{l \in L} h_d(\mathbf{Y}_l^{I_l}) \leq \alpha$ with probability one. The test which rejects $H_{I'}$ if

$$\max_{l \in L} \max_{i \in I_l} T_{l,i} \geq d_{I'}^\alpha$$

has level α by the following:

$$\begin{aligned} P_{H_{I'}} \left(\max_{l \in L} \max_{i \in I_l} T_{l,i} \geq d_{I'}^\alpha \right) &\leq \sum_{l \in L} P_{H_{I'}} \left(\max_{i \in I_l} T_{l,i} \geq d_{I'}^\alpha \right) \quad \text{by Boole's inequality} \\ &= E_{I'} \left[\sum_{l \in L} h_d(\mathbf{Y}_l^{I_l}) \right] \\ &\leq \alpha. \end{aligned}$$

A multiple comparison procedure can then be formed by following the principle of closure described in Section 2.1. Again suppose the observed test statistics are $t_1 \geq t_2 \geq \dots \geq t_k$, corresponding to hypotheses $H_1, H_2, \dots, H_k$. Using the monotonicity property of the critical values defined by (3), results analogous to those obtained in Section 2.3 are obtained for a shortcut procedure. The final procedure can be described compactly in the following algorithm based on the ordered test statistics and hypotheses.

Step Down Multivariate Permutational Algorithm

- 0 Set $j = 1$.
- 1 If $t_j \geq d_{\{j, \dots, k\}}^\alpha$ then reject H_j and continue with step 2, otherwise stop testing.
- 2 Increment $j \rightarrow j + 1$ and return to step 1 if $j \leq k$, otherwise stop testing.

The MCP based on using the above algorithm will be referred to as the Step Down Multivariate Permutation procedure using Closure, or SDMP-C.

2.5 Combining Multiple Comparison Procedures

The SDMP-C is a procedure to test all k hypotheses, while controlling the FWE. For comparison, we now consider procedures constructed by combining existing procedures that are applicable separately to each set of hypothesis types. For this, assume that there are c hypothesis types and k_l outcomes of interest for hypothesis type l , $l = 1, \dots, c$. There are several approaches for testing the hypotheses of a single type. For any of these, it is rather straightforward to apply a Bonferroni-like adjustment to combine the tests on each hypothesis type into a single MCP over the entire set of $k = \sum k_l$ hypotheses. We shall demonstrate by example.

The simplest procedure for testing the k_l hypotheses of a single given type is the Bonferroni procedure itself, which amounts to assigning adjusted p-values by multiplying the raw p-value by k_l and truncating any values over 1. If we denote the Bonferroni adjusted p-values for hypothesis type l by $\tilde{P}_{l,i}$, $l = 1, \dots, c$, $i = 1, \dots, k_l$, then we can combine these values into p-values adjusted for the complete set of k hypotheses, $\tilde{P}_{l,i}$, by multiplying each $\tilde{P}_{l,i}$ further by k/k_l . In this simple case, the resulting MCP is the same as if one had just applied a

Bonferroni correction on all k hypotheses from the start, but in general this will not be the case.

A second possibility for testing the k_l hypotheses of a given type is to use the single step permutational procedure given in Westfall and Troendle (2008). We shall again use the notation that the p-values adjusted for the hypotheses of type l are given by $\dot{P}_{l,i}$. Then we can combine these values into p-values adjusted for the complete set of k hypotheses, $\check{P}_{l,i}$, by multiplying each $\dot{P}_{l,i}$ by k/k_l . We shall refer to the resulting MCP as the Single Step Multivariate Permutation procedure followed by Bonferroni, or SSMP-B.

A third possibility for testing the k_l hypotheses of a given type is to use the step down permutational procedure given in Westfall and Troendle (2008). We shall once again use the notation that the p-values adjusted for the hypotheses of type l are given by $\dot{P}_{l,i}$. Then we can combine these values into p-values adjusted for the complete set of k hypotheses, $\check{P}_{l,i}$, by multiplying each $\dot{P}_{l,i}$ by k/k_l . We shall refer to the resulting MCP as the Step Down Multivariate Permutation procedure followed by Bonferroni, or SDMP-B.

Another possibility for construction of procedures to test all k hypotheses is to extend procedures that test the c hypothesis types on a single outcome variable by applying a Bonferroni-like adjustment for the number of outcome variables. The natural competitor for our permutational procedures is the permutational procedure of Petrondas and Gabriel (1983). For a given outcome variable, this procedure uses closure (and any logical restrictions between the hypotheses) to test the various group hypotheses. Assume that there are q outcome variables and r_i hypotheses of interest for outcome i , $i = 1, \dots, q$. Here we use the notation that the p-values adjusted for the hypotheses on outcome i are given by $\dot{P}_{l,i}$. Then we can combine these values into p-values adjusted for the complete set of k hypotheses, $\check{P}_{l,i}$, by multiplying each $\dot{P}_{l,i}$ by $k=r_i$. We shall refer to the resulting MCP as the Permutation procedure of Petrondas and Gabriel followed by Bonferroni, or PG-B.

For a single outcome variable, we would expect (for most base test statistic choices) the PG-B to be more powerful than the SDMP-C, because of the incorporation of logical restrictions. However, the subsequent application of a Bonferroni-like adjustment for the number of outcome variables leaves the power of the PG-B in doubt. We explore the power of the procedures in Section 4.

3 Adverse Events in a Clinical Trial

The five MCP's described in Section 2 are illustrated by applying them to data on adverse events in a clinical trial. The trial had 160 patients divided into two treatments at two possible dose levels for each treatment. There were four treatment groups in all: low dose of treatment A, high dose of treatment A, low dose of treatment B, high dose of treatment B. Each patient was either observed to have or not, each of 28 adverse events. The data are therefore multivariate Bernoulli. The null hypotheses of interest on adverse event i are (1) no A-dose effect: $F_1^{(i)}=F_2^{(i)}$, (2) no B-dose effect: $F_3^{(i)}=F_4^{(i)}$, (3) no treatment effect: $F_1^{(i)}=F_2^{(i)}=F_3^{(i)}=F_4^{(i)}$. The raw p-values for the base tests for each null hypothesis was obtained by using the standard chi-squared test of independence for the $2 \times g$ table formed by binary counts of success or failure over the g groups.

The procedures yield p-values adjusted for $28 \times 3 = 84$ hypotheses. All but one of the adverse events result in adjusted p-values of 1.0 because of the relatively large number of hypotheses tested. The results for the adverse event with adjusted p-values less than 1.0 are given in Table I. Using the Bonferroni procedure, there does not appear to be a strong case against any of the null hypotheses. However, using a multivariate permutation adjustment

yields more evidence against the null hypothesis that event # 1 has no treatment effect. This evidence is further increased if one uses the closure-based method, SDMP-C.

Although none of the procedures yield adjusted p-values smaller than the usually adopted level for significance, this example illustrates the potential gain by incorporating correlation in the outcomes while using closure to combine different types of hypotheses. Consideration of a different collection of hypotheses or with different data would yield different results. Moreover, here the truth is not known. Therefore the next section is devoted to investigating the properties of the various procedures in cases where the truth is known.

4 Simulations

With the adverse events in a clinical trial example as a template, we designed a simulation experiment to compare the FWE control and average power of the MCP's described in Section 2. The data were generated as correlated multivariate Bernoulli vectors of dimension three, from each of four treatment groups. There were 50 subjects per group, and the null hypotheses were exactly those described in Section 3. The Bernoulli probability in group j is denoted p_j . Correlated Bernoulli vectors were obtained by dichotomizing multivariate Gaussian vectors, generated with correlation chosen so that the resulting Bernoulli vector would have the desired correlation. Only equal correlation is reported, as no important differences were observed using other correlation patterns. The target FWE was set at 5%.

Results of estimated FWE are given in Table II for cases where all of the null hypotheses are true. In the simulations reported here, the FWE is always highest in the complete null case, and therefore is only reported for this case. All of the procedures appear to control the FWE, although they can be quite conservative. The procedures are most conservative in the cases where the success probabilities are small, which causes small number of occurrences and extreme discreteness of the permutation distributions. This problem is well known in the simple testing case to make the Fisher exact test more conservative for smaller success probabilities (Agresti, 2007, Sec. 2.6.3). Small sample modifications to lessen the conservativeness of the Fisher exact test are available for the simple testing case; similar modifications may be developed for multiple testing but are not explored here.

Results of estimated power, averaged over the false null hypotheses, are given in Table III for cases where some or all of the null hypotheses are false. The results show that the power of the Bonferroni procedure can be improved by incorporating correlation through permutations within each hypothesis type, along with a step down testing approach (SDMP-B). Furthermore, this power is similar to that of using closure on each outcome variable separately and then adjusted by the Bonferroni for the different outcome variables (PG-B). Moreover, the SDMP-C, which uses a shortcut version of closure over all of the outcomes and types, improves the power even further over the SDMP-B in cases where each of the null hypotheses are false.

A second simulation experiment was designed based on the three group pairwise comparison problem. The data were again generated as correlated multivariate Bernoulli vectors of dimension three, from each of three treatment groups. There were 50 subjects per group, and the null hypotheses were those of a pair of groups having equal probabilities of success on the i 'th outcome. The target FWE was set at 5%.

Results of estimated FWE are given in Table IV for cases where all of the null hypotheses are true. In the simulations reported here, the FWE is always highest in the complete null case, and therefore is only reported for this case. All of the procedures appear to control the FWE, except the Bonferroni procedure based on the chi-squared p-values does not in one instance. This illustrates one of the advantages of using permutational adjustments for

multiplicity: the resulting procedures are valid regardless of choice of the base tests. In this case the chi-squared tests do not have level control for the small sample size considered here.

Results of estimated power, averaged over the false null hypotheses, are given in Table V for cases where some or all of the null hypotheses are false. The results show much the same pattern as those of the previous experiment with four groups. However, this experiment is an example where the PG-B method gets to take advantage of the logical restrictions between the null hypotheses (it can't happen that exactly two of the three null hypotheses are true), but the shortcut method, SDMP-C, does not. Even in this case, the SDMP-C seems to be preferable.

We also investigated the case of two blocks of five correlated outcomes, where the correlation is zero between blocks and equal within blocks. In either the four group or three group experiments (results not shown) we found no important differences from those reported for equal correlation.

5 Discussion

We have shown that powerful procedures can be constructed that simultaneously account for all of the hypotheses under consideration in a general multivariate multiple group setting. The procedures implicitly account for discreteness, making them more powerful than Bonferroni-type procedures. If a joint distributional assumption can be made, then multivariate permutations can be used and the resulting procedure will also account for correlation between the outcome variables. The procedures are based on applying closure to the tests of intersection hypotheses obtained by applying Boole's inequality to the relevant permutation tests.

In cases where the hypotheses are restricted, the method can be improved further by using full closure, which in general requires evaluation of $O(2^k)$ tests. Such a method appears infeasible for applications such as the adverse events example where $k = 84$, as 2^{84} is computationally prohibitive. However in cases with a small number of tests, the methods shown here can be uniformly improved by applying the permutation tests we have presented to the set of hypotheses in the closure. Westfall and Tobias (2007) provide computationally efficient algorithms for identifying and testing a collection of closed hypotheses in the restricted case, under standard parametric assumptions. Extension to the nonparametric case is a subject for future research.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

References

1. Agresti, A. An Introduction to Categorical Data Analysis. 2. Wiley-Interscience; Hoboken, NJ: 2007.
2. 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.
3. Marcus R, Peritz E, Gabriel KR. On Closed Testing Procedures With Special Reference to Ordered Analysis of Variance. *Biometrika*. 1976; 63:655–660.
4. Petronidas DA, Gabriel KR. Multiple Comparisons by Rerandomization Tests. *Journal of the American Statistical Association*. 1983; 78:949–957.

5. Strasser H, Weber C. On the Asymptotic Theory of Permutation Statistics. *Mathematical Methods of Statistics*. 1999; 8:220–250.
6. Troendle JF. A Stepwise Resampling Method of Multiple Hypothesis Testing. *Journal of the American Statistical Association*. 1995; 90:370–378.
7. Troendle JF. A Permutational Step-Up Method of Testing Multiple Outcomes. *Biometrics*. 1996; 52:846–859. [PubMed: 8805759]
8. 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.
9. Westfall PH, Troendle JF. Multiple Testing With Minimal Assumptions. *Biometrical Journal*. 2008; 50:745–755. [PubMed: 18932134]
10. Westfall, PH.; Young, SS. *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment*. Wiley; New York: 1993.

Table 1

Adjusted p-values for the adverse events in a clinical trial example.

Event #	Hyp. Type	Raw p-value	Bonf.	SSMP-B*	SDMP-B*	SDMP-C*	PG-B*
1	3	0.01113	0.935	0.241	0.241	0.157	0.289

* Permutational adjustments based on 999959 random permutations.

Table II

FWE (%) of the tests for the four group problem under complete null configurations.*

<i>q</i>	<i>p</i> ₁	<i>p</i> ₂	<i>p</i> ₃	<i>p</i> ₄	Corr.	Bonf.	SSMP-B [†]	SDMP-B [†]	SDMP-C [†]	PG-B [†]
3	.5	.5	.5	.5	0.0	4.51	3.35	3.35	3.53	4.70
		.25	.25	.25	0.5	4.17	3.30	3.30	3.49	4.41
			.25	.25	0.7	3.78	3.39	3.39	3.57	4.03
	.25	.25	.25	.25	0.0	3.93	3.41	3.41	3.63	4.76
10					0.5	3.69	3.30	3.30	3.58	4.39
					0.7	3.36	3.44	3.44	3.65	3.96
	.1	.1	.1	.1	0.0	2.75	3.03	3.03	3.55	4.62
		.25	.25	.25	0.5	2.62	3.05	3.05	3.60	4.36
10	.5	.5	.5	.5	0.0	4.16	3.73	3.73	3.76	3.97
					0.5	3.45	3.57	3.57	3.62	3.34
					0.7	2.58	3.54	3.54	3.45	2.64
	.25	.25	.25	.25	0.0	3.62	3.80	3.80	3.77	4.08
10					0.5	3.03	3.67	3.67	3.67	3.40
					0.7	2.40	3.65	3.65	3.66	2.66
	.1	.1	.1	.1	0.0	2.63	3.51	3.51	3.77	3.73
		.25	.25	.25	0.5	2.26	3.38	3.38	3.67	3.01
10					0.7	1.77	3.30	3.30	3.70	2.40

* Estimated from 100000 replications.

[†] Permutational tests based on 959 random permutations.

Table III

Average power (%) of the tests for the four group problem.*

<i>q</i>	<i>p</i> ₁	<i>p</i> ₂	<i>p</i> ₃	<i>p</i> ₄	Corr.	Bonf.	SSMP-B [†]	SDMP-B [†]	SDMP-C [†]	PG-B [†]
3	.8	.5	.5	.5	0.0	68.1	66.8	73.9	73.2	73.7
					0.5	67.9	67.5	73.0	73.1	73.5
					0.7	68.0	68.9	73.3	73.9	73.6
	.8	.5	.2	.5	0.0	77.2	75.3	79.3	86.7	80.5
					0.5	77.1	75.9	79.1	86.4	80.5
					0.7	77.2	76.9	79.4	86.7	80.5
	.8	.5	.8	.5	0.0	73.4	71.6	77.1	83.9	78.3
					0.5	73.4	72.3	76.7	83.3	78.3
10	.8	.5	.5	.5	0.0	73.4	73.4	76.9	83.5	78.3
					0.5	53.5	53.0	63.2	58.6	55.3
					0.5	53.4	55.2	64.0	60.6	55.3
					0.7	53.4	58.6	65.7	63.7	55.3
	.8	.5	.2	.5	0.0	67.6	66.6	72.4	78.3	67.1
					0.5	67.6	68.2	73.4	79.7	67.1
					0.7	67.6	70.5	74.7	81.4	67.1
	.8	.5	.8	.5	0.0	61.1	60.2	69.6	73.1	62.2
					0.5	61.0	62.2	69.9	74.2	62.1
					0.7	61.0	65.1	71.1	76.3	62.1

* Estimated from 100000 replications.

[†] Permutational tests based on 959 random permutations.

Table IV
FWE (%) of the tests for the three group problem under complete null configurations.*

<i>q</i>	<i>p</i> ₁	<i>p</i> ₂	<i>p</i> ₃	Corr.	Bonf.	SSMP-B [†]	SDMP-B [†]	SDMP-C [†]	PG-B [†]
3	.5	.5	.5	0.0	5.41	3.60	3.60	3.79	4.45
				0.5	5.08	3.56	3.56	3.79	4.15
				0.7	4.40	3.56	3.56	3.74	3.71
	.25	.25	.25	0.0	4.39	3.55	3.55	3.98	4.45
10				0.5	3.94	3.35	3.35	3.78	4.00
				0.7	3.45	3.28	3.28	3.71	3.50
	.1	.1	.1	0.0	2.87	2.91	2.91	3.61	3.62
				0.5	2.57	2.72	2.72	3.46	3.30
10	.5	.5	.5	0.0	4.92	3.98	3.98	3.96	3.19
				0.5	4.17	3.95	3.95	3.96	2.73
				0.7	3.16	3.94	3.94	3.77	2.13
	.25	.25	.25	0.0	3.99	4.06	4.06	4.10	3.15
10				0.5	3.27	3.78	3.78	3.87	2.70
				0.7	2.57	3.93	3.93	4.05	2.08
	.1	.1	.1	0.0	2.33	3.51	3.51	3.73	2.07
				0.5	2.18	3.44	3.44	3.88	1.99
10				0.7	1.52	3.11	3.11	3.57	1.40

* Estimated from 100000 replications.

[†] Permutational tests based on 959 random permutations.

Table V

Average power (%) of the tests for the three group problem.*

q	p_1	p_2	p_3	Corr.	Bonf.	SSMP-B [†]	SDMP-B [†]	SDMP-C [†]	PG-B [†]
3	.8	.5	.5	0.0	43.9	42.1	46.0	46.7	46.8
				0.5	43.8	42.6	45.7	46.8	46.7
				0.7	43.8	43.5	46.0	47.4	46.7
	.8	.5	.2	0.0	57.9	56.5	59.4	65.3	61.3
				0.5	57.9	57.0	59.4	65.3	61.3
				0.7	57.9	57.7	59.5	65.6	61.3
10	.8	.5	.5	0.0	34.4	33.3	39.1	37.1	33.2
				0.5	34.4	34.9	40.0	38.5	33.1
				0.7	34.2	37.1	41.1	40.4	33.0
	.8	.5	.2	0.0	50.8	50.0	54.4	58.8	51.8
				0.5	50.8	51.3	55.1	60.1	51.8
				0.7	50.8	52.9	56.0	61.6	51.7

* Estimated from 100000 replications.

[†] Permutational tests based on 959 random permutations.