A Note on Controlling the Number of False Positives

Edward L. Korn* and Boris Freidlin

Biometric Research Branch, National Cancer Institute, Bethesda, Maryland 20852-7434, U.S.A.
*email: korne@ctep.nci.nih.gov

Summary. Lehmann and Romano (2005, Annals of Statistics 33, 1138–1154) discuss a Bonferroni-type procedure that bounds the probability that the number of false positives is larger than a specified number. We note that this procedure will have poor power as compared to a multivariate permutation test type procedure when the experimental design accommodates a permutation test. An example is given involving gene expression microarray data of breast cancer tumors.

Key words: Bonferroni; False discovery proportion; Multiple testing; Multivariate permutation test; Permutation test.

1. Introduction

In many applications involving high-dimensional data, there may be thousands of null hypotheses being considered (e.g., one for each of 10,000 genes), and there may be interest in identifying which null hypotheses, if any, can be rejected. Controlling the family-wise error rate is typically too stringent a criterion, as it may lead to limited power to detect real differences. In these cases it may be useful to limit the number of false positives or the proportion of false positives (false discovery proportion) with high probability (Korn et al., 2004). (An alternative is to estimate the expected false discovery proportion, which is sometimes known as the false discovery rate [Benjamini and Hochberg, 1995]; see Ge, Dudoit, and Speed [2003] and Li et al. [2005] for extensive reviews of the false discovery rate and Korn et al. [2004] for a discussion of the differences between estimating the false discovery proportion and the false discovery rate.) Lehmann and Romano (2005) recently investigated a simple Bonferroni-type procedure (Hommel and Hoffmann, 1988) to limit the type procedure rejects all hypotheses whose p-values are ≤α for all i ∈ {0, 1} (1)

when the ith null hypothesis is true. When the null hypotheses of interest can be tested with a permutation test, an alternative approach is based on the reference distribution of an appropriate order statistic of the univariate p-values using the multivariate permutation distribution; see Westfall and Young (1993), Korn et al. (2004), Romano and Wolf (2007), and Section 2 below.

In this note we demonstrate that in applications where a multivariate permutation test (MPT)-type procedure is appropriate, the Bonferroni-type procedure would be expected to have poor power as compared to the MPT-type procedure. We focus on the two-class comparison, a simple situation in which the MPT-type procedure is performed by permuting the class labels between the two classes. For simplicity of presentation, we consider only single-step procedures in this article; in applications where only a very small proportion of nonnull hypotheses are expected, the gains of using a multistep procedure are expected to be small. For example, with s hypotheses, the nominal p-value cut-offs for the Bonferroni procedure (α/s) and the Holm (Holm, 1979) procedure (α/(s − i + 1) for the ith hypothesis tested) are close when s ≫ i. Also for simplicity, we only consider controlling the number of false positives, as procedures that have poor power for this objective would be expected to have poor power for controlling the false discovery proportion. In particular, a procedure for controlling the number of false positives can be modified to control approximately the false discovery proportion by estimating the false discovery proportion as the number of allowable false positives divided by the number of rejected null hypothesis; see Korn et al. (2007).

2. Two-Class Comparison

Let x1, ..., xn and y1, ..., yn be s-dimensional vectors for the observations in class 1 and 2, respectively. The Bonferroni-type procedure uses a univariate p-value (associated with testing the null hypothesis for that variable) for each of the s variables. For example, the null hypothesis for the kth variable may be that the two class means are the same, and the p-value may be from a two-sample t-statistic. To ensure that k or more false positives occur with ≤α probability, the Bonferroni-type procedure rejects all hypotheses whose p-values are ≤kα/s. Letting sk be the number of the s hypotheses that are rejected, and N be the number of these sk null hypotheses that are rejected, the desired probability constraint follows from (Hommel and Hoffmann, 1988; Lehmann and Romano, 2005)

P(N ≥ k) = E(N)k − k−1 i=1 iP(N = i) + k−1 i=1 iP(N = k + i)

≤ E(N)k ≤ α/Nk ≤ α,

(2)
where $\alpha_* \equiv k\alpha/s$ and the validity (1) of the $p$-values is used for the penultimate inequality.

For the MPT procedure to ensure that $k$ or more false positives occur with $\leq \alpha$ probability, let $W_i$ be a univariate statistic associated with the $i$th variable such that smaller values of $W_i$ suggest the $i$th null hypothesis is not true, $i = 1, \ldots, s$. (For the two-class comparison, we assume that the $s$-dimensional multivariate distribution of the variables associated with null hypotheses is the same regardless of class label.) For example, $W_i$ could be a $p$-value associated with a hypothesis test, but need not be. Consider constructing permuted datasets by permuting the class labels, and, for each permuted dataset, calculating the $W$'s on all the variables, say, $W_1^*, \ldots, W_s^*$. Let the ordered $W$'s be denoted $W_{(1)} \leq \cdots \leq W_{(s)}$ and let

$$c_{\alpha,k} = \text{MAX} \left\{ c \mid P\{W_{(k)} < c \mid \{x_1, \ldots, x_n, y_1, \ldots, y_m\}\} \leq \alpha \right\},$$

(3)

where $P\{\cdot \mid \{x_1, \ldots, x_n, y_1, \ldots, y_m\}\}$ refers to the probability under the permutation distribution, and the MAX is the maximum over $c$. (The quantity $c_{\alpha,k}$ is essentially the $k$th quantile of $W_{(k)}$.) The MPT-based procedure rejects all null hypotheses associated with variables $i$ such that $W_i < c_{\alpha,k}$.

To see why this procedure satisfies the probability error constraints, let $I \subseteq \{1, \ldots, s\}$ be the set of indices corresponding to true null hypotheses, let $W_{(i)}^0 \leq \cdots \leq W_{(s)}^0$ be the ordered $W$ statistics on the original (unpermuted) dataset restricted to $i \in I$, let $W_{(i)}^\circ \leq \cdots \leq W_{(s)}^\circ$ be the ordered $W$ statistics on a permuted dataset restricted to $i \in I$, and let

$$c_{\alpha,k} = \text{MAX} \left\{ c \mid P\{W_{(k)}^\circ < c \mid \{x_1, \ldots, x_n, y_1, \ldots, y_m\}\} \leq \alpha \right\}.$$

Note that although $c_{\alpha,k}^\circ$ is unknown to us, we do know that $c_{\alpha,k}^\circ \geq c_{\alpha,k}$ because $W_{(k)}^0 \geq W_{(k)}$ (the $\{W^\circ\}$'s being a subset of the $\{W^*\}$). The proof that the probability that $k$ or more null hypotheses are rejected is $\leq \alpha$ is as follows:

$$P(k \text{ or more null hypotheses rejected})$$

$$= P(W_{(k)}^0 < c_{\alpha,k})$$

$$= E\left[P(W_{(k)}^\circ < c_{\alpha,k} \mid \{X_1, \ldots, X_n, Y_1, \ldots, Y_m\}\right].$$

$$\leq E\left[P(W_{(k)}^\circ < c_{\alpha,k}^\circ \mid \{X_1, \ldots, X_n, Y_1, \ldots, Y_m\}\right].$$

$$\leq E(\alpha) = \alpha.$$

For our comparison of the Bonferroni-type procedure with the MPT-based procedure, we will assume that a $p$-value from a two-sample $t$-test is used for each variable for the Bonferroni-type procedure, say $p_i$, and these same $p$-values as the test statistics for the MPT-based procedure, i.e., $W_i = p_i$, $i = 1, \ldots, s$. First, consider the case when $n, m \to \infty$ (asymptotic sample size case) and the global null hypothesis that $\{X_1, \ldots, X_n, Y_1, \ldots, Y_m\}$ independent and identically distributed from a $s$-dimensional continuous distribution that has correlation matrix $R$ and whose components have finite absolute third moments. Let $F_{n,m}(u_1, \ldots, u_s \mid \{X_1, \ldots, X_n, Y_1, \ldots, Y_m\})$ be the cumulative distribution function of the permutation distribution of the vector of $t$-statistics. Following Hoeffding (1952), $F_{n,m}(u_1, \ldots, u_s \mid \{X_1, \ldots, X_n, Y_1, \ldots, Y_m\})$ converges in probability to the cumulative distribution function of a multivariate normal distribution with mean 0 and covariance (and correlation) matrix $R$. The following theorem, whose proof is given in the Appendix, shows that (i) we can use a multivariate normal distribution to evaluate the asymptotic properties of the MPT-based procedure and that (ii) the asymptotic $p$-value cut-off for rejection for the MPT-based procedure is greater than or equal Bonferroni-type procedure.

**Theorem 1:** Let $\{X_1, \ldots, X_n, Y_1, \ldots, Y_m\}$ be independent and identically distributed from an $s$-dimensional continuous distribution that has correlation matrix $R$ and whose components have finite absolute third moments. Let $c_{\alpha,k}^{(n,m)}(X_1, \ldots, X_n, Y_1, \ldots, Y_m)$ be the $c_{\alpha,k}$ defined by (3) with the explicit notation for the sample sizes and conditioning order statistics.

(i) \[
\lim_{n,m \to \infty} c_{\alpha,k}^{(n,m)}(X_1, \ldots, X_n, Y_1, \ldots, Y_m) \to c_{\alpha,k}^{(\infty)} = 2(1 - \Phi(\gamma_{\alpha,k})),
\]

where $\gamma_{\alpha,k}$ is the $1 - \alpha$ quantile of the $k$th largest of $|Z_1|, \ldots, |Z_s|$, and where $Z = (Z_1, \ldots, Z_s)$ has a multivariate normal distribution with mean 0 and covariance matrix $R$, and $\Phi(*)$ is a standard normal cumulative distribution function.

(ii) $c_{\alpha,k}^{\infty} \geq \frac{k\alpha}{s}$.

To see how much better the power is for the MPT-based procedure as compared to the Bonferroni-type procedure, we conducted a simulation that would be relevant to microarray experiments: 10,000 variables with a block diagonal correlation structure with block size of 100 and correlation $p$ within the blocks. For the asymptotic sample size case, Table 1 displays the cut-offs required for the two procedures ($\alpha = 0.05$) to reject the null hypothesis for any given variable in terms of the nominal $p$-value for that variable. For the Bonferroni-type

### Table 1

<table>
<thead>
<tr>
<th>$\rho$</th>
<th>Number of MPT-based errors ($k - 1$)</th>
<th>Bonferroni-type procedure (MPT-based procedure ($c_{\alpha,k}^{(n,m)}$))</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.5</td>
<td>0.51 0.61</td>
</tr>
<tr>
<td>1</td>
<td>1.0</td>
<td>3.53 2.90</td>
</tr>
<tr>
<td>2</td>
<td>1.5</td>
<td>8.14 5.86</td>
</tr>
<tr>
<td>3</td>
<td>2.0</td>
<td>13.61 9.18</td>
</tr>
<tr>
<td>4</td>
<td>2.5</td>
<td>19.70 12.72</td>
</tr>
<tr>
<td>5</td>
<td>3.0</td>
<td>26.10 16.45</td>
</tr>
<tr>
<td>6</td>
<td>3.5</td>
<td>32.76 20.34</td>
</tr>
<tr>
<td>7</td>
<td>4.0</td>
<td>39.71 24.36</td>
</tr>
<tr>
<td>8</td>
<td>4.5</td>
<td>46.99 28.53</td>
</tr>
<tr>
<td>9</td>
<td>5.0</td>
<td>54.28 32.77</td>
</tr>
<tr>
<td>10</td>
<td>5.5</td>
<td>61.73 37.13</td>
</tr>
</tbody>
</table>
Note on Controlling the Number of False Positives

229

Table 2

<table>
<thead>
<tr>
<th>Number of allowable errors (k = 1)</th>
<th>Bonferroni-type procedure (%)</th>
<th>MPT-based procedure</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ρ = 0 (%)</td>
<td>ρ = 0.5 (%)</td>
</tr>
<tr>
<td>0</td>
<td>67</td>
<td>68</td>
</tr>
<tr>
<td>1</td>
<td>72</td>
<td>79</td>
</tr>
<tr>
<td>2</td>
<td>75</td>
<td>84</td>
</tr>
<tr>
<td>3</td>
<td>77</td>
<td>86</td>
</tr>
<tr>
<td>4</td>
<td>78</td>
<td>88</td>
</tr>
<tr>
<td>5</td>
<td>80</td>
<td>89</td>
</tr>
<tr>
<td>6</td>
<td>81</td>
<td>90</td>
</tr>
<tr>
<td>7</td>
<td>81</td>
<td>91</td>
</tr>
<tr>
<td>8</td>
<td>82</td>
<td>91</td>
</tr>
<tr>
<td>9</td>
<td>83</td>
<td>92</td>
</tr>
<tr>
<td>10</td>
<td>83</td>
<td>93</td>
</tr>
</tbody>
</table>

and for k = 2

\[ P(N \geq 2) = \frac{\alpha^k s}{2} - \frac{\sum_{i=1}^{s-2} i P(N = 2 + i)}{2} \]

\[ P(N = 1) + \sum_{i=1}^{s-2} i P(N = 2 + i) \]

\[ \alpha - \frac{1}{2} \sum_{i=1}^{s-1} \frac{(1 - \alpha^k)^{i-1}}{i} \approx \alpha - 2\alpha^3. \]

The difference between α and the type 1 error of the Bonferroni procedure (k = 1) is approximately

\[ P(N = 2) = \frac{s(s-1)}{2} (\alpha^k)^s (1 - \alpha^k)^{s-2} \approx \alpha^2/2, \]

whereas for the Bonferroni-type procedure with k = 2 it is approximately

\[ \frac{1}{2} P(N = 1) = \frac{1}{2} s \alpha^k (1 - \alpha^k)^{s-1} \approx \alpha - 2\alpha^3. \]

Compared to the target level of α, we see that the conservativeness of the Bonferroni procedure (k = 1) is small (α^2/2), but the conservativeness of the Bonferroni-type procedure (k = 2) is large (α - 2α^3). This conservativeness of the level under the global null hypothesis translates into reduced power when there are nonnull hypotheses.

As the results in Tables 1 and 2 are asymptotic, we also consider a small-sample situation with sample sizes n = 10 and m = 5 in the two groups. A block correlation structure with ρ = 0.5 is again used, with 190 of the normally distributed 10,000 variables nonnull with a true mean difference of 7.9 standard errors (a difference of 7.9 standard errors corresponds to approximately 65% power for a Bonferroni-adjusted t-test with 13 degrees of freedom); 10 of the blocks have 10 nonnull variables each, and the other 90 blocks each have one nonnull variable. Table 3 displays power to reject the null hypothesis for a nonnull variable and the quantities of the distribution of the number of the 190 nonnull hypotheses rejected by the Bonferroni-type and MPT-based procedures. Allowing for zero errors (k = 1), the Bonferroni and MPT-based procedures have similar characteristics. When more than zero errors are allowed, the Bonferroni-type procedure results in a substantial loss of power relative to the MPT-based procedure. This loss of power seen in Table 3 is even more dramatic than the asymptotic results given in Table 2, e.g., 75% versus 88% for k = 2 (Table 3) and 72% versus 79% for k = 2 (Table 2). The advantage of the MPT-based procedure is also seen by the marked increase in the number of the rejected nonnull hypothesis, e.g., median 167 versus 142 for k = 2. Small-sample properties of the methods are discussed additionally below.

3. An Example

Sotiriou et al. (2003) analyzed cDNA gene expression profiles from 99 tumor specimens from breast cancer patients. In addition to gene expression values for 7650 genes (probes) preprocessed as described in Sotiriou et al. (2003), there was standard prognostic variable information available for each patient. (The data are publicly available at http://linus.nci.nih.gov/~brb/DataArchive.html.) Here we consider
two two-class comparisons based on parametric two-sample t-tests and control for the number of false positive at the \( \alpha \leq 0.05 \) level. For each comparison, we restrict attention to genes for which the number of missing values was less than the number of specimens in the class with fewer observations minus 2.

The first comparison is for patients with grade 1 or 2 tumors \( (n = 54) \) versus patients with grade 3 tumors \( (m = 45) \) with \( s = 7498 \) genes. Allowing for no errors \( (k = 1) \), Bonferroni identifies three genes and the MPT-based procedure identifies six genes. However, allowing for 10 errors \( (k = 11) \), the Bonferroni-type procedure identifies 28 genes and the MPT-based method identifies 94 genes. An interesting gene found by the MPT-based procedure and not by the Bonferroni-type procedure is BUB1, whose gene expression has been previously associated with survival in breast cancer patients (Glinsky, Berezovska, and Glinskii, 2005).

The second comparison is for patients with estrogen receptor negative status \( (n = 34) \) versus patients with estrogen receptor positive status \( (m = 65) \) with \( s = 7470 \) genes. Allowing for no errors \( (k = 1) \), Bonferroni identifies 163 genes and the MPT-based procedure identifies 172 genes. However, allowing for 10 errors \( (k = 11) \), the Bonferroni-type procedure identifies 290 genes and the MPT-based method identifies 503 genes. Two interesting genes found in the 503 gene set but not in the 290 gene set are TSG101 and BAP1. TSG101 represses transcriptional activation by estrogen receptor (Sun et al., 2009), and BAP1 may be a breast cancer tumor suppressor gene (Jensen et al., 1998).

### 4. Discussion

In addition to the lower power as compared to the MPT-based procedure, a severe limitation of the Bonferroni-type method is that it requires a valid univariate p-value for each hypothesis test in the sense of (1). In particular, for p-values derived from parametric tests this is a challenge for very small \( u \) (Ringwald, 1983). For example, a small amount of nonnormality in the \( x \) and \( y \) data will lead to a violation of (1) for very small \( u \) for a t-test derived p-value unless the sample sizes are extremely large. In particular, if a normal critical value corresponding to a \( 10^{-5} \) two-sided p-value was incorrectly used instead of a t-distribution with 100 degrees of freedom, then the actual rejection probability would be \( 2.5 \times 10^{-4} \), far from the nominal level. An approach to avoid this problem is to use p-values from univariate permutation tests, e.g., the Wilcoxon-rank sum test for the two-class comparison. Although these p-values will satisfy (1), they unfortunately lead to a Bonferroni-type procedure with very little power unless the sample sizes are moderately large. For example, for a two-class comparison with \( n = 10 \) and \( m = 5 \), the smallest (two-sided) p-value obtainable from a rank test is \( 6.66 \times 10^{-4} \), leading to no possible rejection for reasonably sized \( k \) and \( s = 10,000 \) variables.

The MPT-based procedure does not require (1) to hold for the procedure to be valid. In fact, any univariate statistic can be used, not necessarily a p-value. The choice of the statistic can obviously affect the power of the procedure, so we generally recommend using p-values from parametric tests. The MPT-based procedure is restricted to experimental designs that accommodate a permutation test. Fortunately, this covers many practical applications: paired or unpaired two-class comparisons, k-class comparisons, simple linear regression, simple logistic regression, and survival analysis with one independent variable. For situations in which a permutation test cannot be directly used, e.g., testing one independent variable in a multiple linear regression, one might attempt to obtain the required reference distribution by bootstrapping the data vectors rather than permuting them under a null model. Unfortunately, the properties of bootstraps with high-dimensional data are unpredictable unless the sample sizes are extremely large (Troendle, Korn, and Meshane, 2004). We believe a more promising approach is to use an approximate permutation test based on permuting the residuals; this is an area of further research.

### Table 3

<table>
<thead>
<tr>
<th>Number of allowable errors ( (k - 1) )</th>
<th>Bonferroni-type procedure</th>
<th>MPT-based procedure</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Power(^a)</td>
<td>First quartile</td>
</tr>
<tr>
<td>0</td>
<td>0.655</td>
<td>117</td>
</tr>
<tr>
<td>1</td>
<td>0.747</td>
<td>137</td>
</tr>
<tr>
<td>2</td>
<td>0.800</td>
<td>147</td>
</tr>
<tr>
<td>3</td>
<td>0.839</td>
<td>153</td>
</tr>
<tr>
<td>4</td>
<td>0.858</td>
<td>158</td>
</tr>
<tr>
<td>5</td>
<td>0.875</td>
<td>162</td>
</tr>
<tr>
<td>6</td>
<td>0.886</td>
<td>165</td>
</tr>
</tbody>
</table>

\(^a\)Power to reject the null hypothesis for a nonnull variable.
ACKNOWLEDGEMENTS

The authors wish to thank Dr Lara Lusa (Department of Experimental Oncology, Fondazione IRCCS Istituto Nazionale dei Tumori, Milan, Italy) for her help in interpreting the results of the example given in Section 3.

REFERENCES


Received December 2006. Revised May 2007. Accepted May 2007.

APPENDIX

Proof of the Theorem 1

(i) The cumulative distribution function of an order statistic can be written in terms of a linear combination of the associated multivariate cumulative distribution function. Therefore, in particular, \( c_{n,k}^{(n,m)}(X_1, \ldots, X_n, Y_1, \ldots, Y_m) \) is the maximum (over \( c \)) for which \( 1 - \alpha \) is less than a linear combination of \( F_{n,m}(u_1, \ldots, u_{t}) \) for various values of \( (u_1, \ldots, u_{t}) \) that are each equal to \( t_{n+m-2}^{-1}(1-c/2) \) or \( \infty \), where \( t_{n+m-2}(\cdot) \) is the inverse cumulative distribution function of a \( t \) distribution with \( n+m-2 \) degrees of freedom. For example, for \( k = 1 \),

\[
\begin{align*}
&c_{\alpha,1}^{(n,m)}(X_1, \ldots, X_n, Y_1, \ldots, Y_m) \\
&= \text{MAX} \{ c | F_{n,m}(t_{n+m-2}^{-1}(1-c/2), \ldots, t_{n+m-2}^{-1}(1-c/2)) \\
&\quad (X_1, \ldots, X_n, Y_1, \ldots, Y_m) \} > 1 - \alpha \}
\end{align*}
\]

The result follows by taking the limit as \( n, m \to \infty \).

(ii) Let \( (Z_1, \ldots, Z_s) \) have a multivariate normal distribution with mean \( 0 \) and covariance matrix \( R \), and let \( p_{(k)} \) be the \( k \)-th smallest value of \( 2|1 - \Phi(Z_i)| \), \( i = 1, \ldots, s \). Then, retesting (2), \( P(p_{(k)} < \frac{c_{n,m}}{s}) \leq \alpha \) (for any \( n \) and \( m \)). Taking the limit of this quantity, one has \( \lim_{n,m \to \infty} P(p_{(k)} < \frac{c_{n,m}}{s}) = d \leq \alpha \). Because the distribution of the \( p_{(k)} \) is continuous, the result follows by contradiction by noting \( \lim_{n,m \to \infty} P(p_{(k)} < \frac{c_{n,m}}{s}) = \alpha \).