Biostatistics (2008), **9**, 3, *pp.* 458–466
doi:10.1093/biostatistics/kxm047
Advance Access publication on December 18, 2007

Significance levels for studies with correlated test statistics

JIANXIN SHI, DOUGLAS F. LEVINSON

Department of Psychiatry and Behavioral Science,
Stanford University School of Medicine, Stanford, CA 94305, USA

ALICE S. WHITTEMORE*

Department of Health Research and Policy, Redwood Building, Room T204,
Stanford University School of Medicine, Stanford, CA 94305, USA
alicesw@stanford.edu

Summary

When testing large numbers of null hypotheses, one needs to assess the evidence against the global null hypothesis that none of the hypotheses is false. Such evidence typically is based on the test statistic of the largest magnitude, whose statistical significance is evaluated by permuting the sample units to simulate its null distribution. Efron (2007) has noted that correlation among the test statistics can induce substantial interstudy variation in the shapes of their histograms, which may cause misleading tail counts. Here, we show that permutation-based estimates of the overall significance level also can be misleading when the test statistics are correlated. We propose that such estimates be conditioned on a simple measure of the spread of the observed histogram, and we provide a method for obtaining conditional significance levels. We justify this conditioning using the conditionality principle described by Cox and Hinkley (1974). Application of the method to gene expression data illustrates the circumstances when conditional significance levels are needed.

Keywords: Conditional p-value; Gene expression data; Genome-wide association data; Multiple testing; Overall p-value.

1. Introduction

Many studies involve testing large numbers of hypotheses. For example, we may be interested in whether the expression levels of a large number M of genes are similar in 2 types of cancer tissue or whether genotypes of M genetic markers are similar in individuals with and without a given disease. To address such questions, investigators obtain random samples from the 2 populations and form M test statistics $T_1, \ldots, T_M$ evaluating the differences. An important analytic starting point is assessment of the overall evidence against the global null hypothesis H_0 of no differences between the 2 populations, that is, that none of the M null hypotheses is false. Such evidence, when present, motivates the search for specific

*To whom correspondence should be addressed.

attributes that differ in the populations. Even if such evidence is absent, the study can provide important leads for investigation in other studies. Ultimately, however, one is unlikely to declare that the populations differ in any specific attributes unless evidence against H_0 has been observed in several independent studies.

There are several strategies for testing H_0 (see Westfall and Young, 1993, for a review). Here, we shall work with the test statistic having largest magnitude. For analytic convenience, we first convert each test statistic to a z-value:

$$Z_m = \Phi^{-1}[\Psi(T_m)], \quad m = 1, \ldots, M, \tag{1.1}$$

where Φ is the standard normal cumulative distribution function (cdf) and Ψ is a putative null marginal cdf for each statistic. Under H_0, each Z_m has a standard normal marginal distribution. Let $X = \max_m |Z_m|$. If the M statistics were independent, the overall p-value associated with the observed value $X = x$ would be $P_x = 1 - [1 - 2\Phi(-x)]^M \sim 2M\Phi(-x)$, where the approximation holds for large x. When the test statistics are correlated, we can estimate P_x by simulating the global null distribution of X. This is done by permuting the sample units many times. In each permutation, we compute M test statistics and record the value of the one with largest magnitude. We estimate P_x as the proportion of all permutations whose maximum test statistic exceeds the observed value x.

However, there is a potential problem with using P_x as a measure of overall significance. Specifically, the shape of the histogram of values Z_m varies across studies with correlated tests more than it does across studies with independent tests (Owen, 2005; Efron, 2007). The larger variation among histograms of correlated statistics can cause misleading tail counts, and some adjustment is needed to avoid erroneous inferences.

Here, we describe those situations when histogram variability can produce misleading values for an overall significance level. We begin with a more formal description of the problem and our proposal for addressing it. We then review a conditionality principle that describes the circumstances when conditional significance levels are more appropriate than unconditional ones. We present a simple, permutation-based method for estimating the difference between conditional and unconditional significance levels, while also estimating the variance of C (and its covariance with tail counts) for comparison with expected values. The investigator can then consider whether conditional p-values would substantially alter the interpretation of results and report them when appropriate. We also derive approximations that elucidate the 2 key factors determining the need for conditional significance levels. The first factor is the average of the squared correlation coefficients for all pairs of distinct test statistics. The second factor is the magnitude x of the maximum test statistic. Finally, we illustrate the method by application to gene expression data.

2. The problem

Consider the sample histograms of data from several studies, each producing M test statistics, $Z_1, \ldots, Z_M$, with each Z_m having a standard Gaussian marginal distribution. If the Z_m are independent, the histograms vary across studies in predictable ways. For example, we can calculate various moments of the distribution of a histogram's "central proportion" C, that is, the proportion of test statistics Z_m with absolute value less than 1: $C = \#\{Z_m\colon |Z_m| < 1\}/M$. The expected value of C is $\Phi(1) - \Phi(-1) = 0.6827$, and its variance is $0.6827(1 - 0.6827)/M = 0.2166/M$. However, correlation among the test statistics increases the variability of the sample histograms (Efron, 2007). For instance, with $M = 2000$ test statistics, the standard deviation of C increases from 0.01 under independence to about 0.08 when the average squared correlation coefficient between pairs of test statistics is 0.05. Of particular importance, extreme values of the test statistics can be more variable than appreciated in the presence of correlation, which can complicate interpretation of results from any given study. Because a permutation-based estimate of the overall p-value P_x does not account for the excess variability, significance levels assigned to an observed value $X = x$ can be misleading.

2.1 *Conditioning on the central proportion C*

We propose assessing the variability of X via that of C and, when the variability is large, reducing it by conditioning on the observed value $C = c$. That is, we replace the permutation-based unconditional p-value P_x by the conditional p-value

$$p_x(c) = \Pr(X > x|C = c) = E[Y|C = c].$$

Here, $Y = 1$ if $X > x$ and $Y = 0$ otherwise.

This variance-reducing tactic is familiar to statisticians. Our statistic of interest Y is correlated with another statistic C that is not of interest. (C is not of interest because the test statistics between ±1 are unlikely to represent hypotheses that are false and thus "interesting.") Moreover, the variance of C is greater than expected under standard assumptions, such as independence of the test statistics. This is precisely the setup for the conditionality principle described by Cox and Hinkley (1974). The principle states that when C has large variance and is strongly correlated with Y, inferences based on Y should be drawn as if C were fixed at its observed value. Other values that C might have taken are irrelevant to the data at hand.

2.2 *Example illustrating the conditionality principle*

Suppose we randomly choose a coin from a bag containing 90% nickels and 10% quarters. We flip our chosen coin and then use the outcome (heads or tails) to select one of 2 hypotheses. The first hypothesis states that the nickels are biased toward heads, while the second hypothesis states that the quarters are so biased. Columns 1 and 2 of Table 1 show the probabilities of heads specified by the 2 hypotheses for each of the 2 choices of coin. The unconditional probability of heads is $0.9(0.8) + 0.1(0.1) = 0.73$ according to hypothesis 1 and $0.9(0.1) + 0.1(0.8) = 0.17$ according to hypothesis 2. Suppose we choose a quarter, flip it, and get heads. If we ignore the coin type chosen, the probabilities in column 3 would prompt us to favor hypothesis 1. But if we condition on the fact that we chose a quarter, we would favor hypothesis 2. Thus, directly conflicting evidence is provided if we do not condition on the type of coin chosen.

In this example, there is a strong association between the heads/tails outcome of interest and the choice C of coin, and this choice varies from one experiment to another. The likelihood of substantial differences between conditional and unconditional inferences increases with the strength of dependence between the outcome of interest and the coin selection C, and it increases with the variance of C. The example shows that if the covariance between an extreme-value statistic and an appropriately chosen conditioning statistic C is sufficiently large, and if C is sufficiently variable, then different conclusions about H_0 may be drawn, depending on whether or not we condition on C.

Table 1. *Probability of heads when flipping a randomly selected coin, under* 2 *hypotheses*

Hypothesis	Heads probability		
	Conditional on coin type[†]		Unconditional
	Nickel	Quarter	
Nickels biased toward heads	0.80	0.10	0.73
Quarters biased toward heads	0.10	0.80	0.17

[†] Coin is selected from a bag containing 90% nickels and 10% quarters.

2.3 *When are conditional p-values needed?*

The most straightforward way to evaluate the need for conditional p-values is to estimate both conditional and unconditional p-values and compare them. This can be done by repeatedly permuting the sample units and generating M test statistics for each permutation. The unconditional p-value is estimated as the proportion of permutations whose maximal test statistic exceeds x. In principle, the conditional p-value could be estimated by restricting the permutations to those whose C-value lies in a small interval around the observed value $C = c$ and basing the p-value estimate only on the permutation outcomes within that small subset. For example, for a study comparing the M attributes of 2 samples of N_1 and N_2 units, the rth of the R permutations would involve the following steps: (i) randomly select a subset of N_1 of all the units, call it sample 1, and call the remaining N_2 units sample 2; (ii) compute the M test statistics comparing the 2 samples; and (iii) flag the permutation if the magnitude of any test statistic exceeds x. Repeating these 3 steps R times, we estimate the unconditional overall p-value as $\widehat{P_x} = \#\{\text{flagged permutations}\}/R$. In addition, we might estimate the conditional overall p-value, given conditioning statistic $C = c$, as

$$\widehat{p_x}(c) = \frac{\#\{\text{flagged permutations with } C_r \in (c-\delta, c+\delta)\}}{\#\{\text{permutations with } C_r \in (c-\delta, c+\delta)\}}, \tag{2.1}$$

where δ is a small positive number.

A limitation of the latter estimate is the large number of permutations needed to achieve adequate precision, particularly when the observed c falls in the tail of the distribution of C, a circumstance yielding the largest potential difference between conditional and unconditional p-values. Instead, we propose inferring $p_x(c)$ by regression smoothing. Specifically, let $Y_r = 1$ if the test statistic of largest magnitude in the rth permuted data set is flagged, with $Y_r = 0$ otherwise. We fit to the points $\{(Y_r, C_r), r = 1, \ldots, R\}$ the logistic regression model $\text{logit}[\Pr(Y_r = 1|C_r)] = \alpha + \beta C_r$ to obtain fitted parameter values $\hat{\alpha}$ and $\hat{\beta}$. We then use these to estimate the conditional p-value given the observed $C = c$ as

$$\widehat{p_x}(c) = \frac{\exp(\hat{\alpha} + \hat{\beta}c)}{1 + \exp(\hat{\alpha} + \hat{\beta}c)}. \tag{2.2}$$

If the 2 p-values $\widehat{P_x}$ and $\widehat{p_x}(c)$ differ appreciably, then the conditionality principle indicates that the latter is a better summary of the evidence against H_0. This regression smoothing borrows strength from the full ensemble of R permutations instead of just those with C close to c. It is similar in flavor to the regression smoothing of Efron (2007) when inferring the false discovery rate and the extreme-value smoothing of Dudbridge and Koeleman (2004). Using simulated data sets (not shown), we have found that the smoothed conditional p-values (2.2) agree closely with those obtained from extensive permutations using (2.1), for a wide range of values of C and X.

The permutations also can provide a fuller picture of the need for conditioning, by providing empirical estimates of the moments $\text{var}(C)$ and $\text{cov}(C, Y)$. The extent to which these empirical moments exceed their expected values under the independence assumption indicates the need for conditioning. We show in the Appendix that when all pairs of test statistics have bivariate Gaussian distributions, both these moments have approximate representations in terms of the mean of the squared correlation coefficients for all pairs of test statistics,

$$\tau^2 = \frac{2}{M(M-1)} \sum_{1 \leqslant m < m' \leqslant M} \rho_{mm'}^2.$$

Specifically,

$$\text{var}(C) \sim \frac{0.2168}{M} + 0.11713\tau^2 \tag{2.3}$$

and

$$\operatorname{cov}(C, Y) \sim -P_x\left(\frac{0.6826}{M} + 0.2040x^2\tau^2\right). \tag{2.4}$$

These approximations hold for small values of τ^2 and large values of x. Expressions (2.3) and (2.4) show that potential differences between conditional and unconditional significance levels increase with the correlation measure τ^2 and the threshold value x.

3. APPLICATION TO GENE EXPRESSION DATA

Here, we apply the method to expression levels of $M = 3226$ genes in the breast cancers of 15 patients with pathogenic mutations of the tumor-suppressor genes BRCA1 ($N_1 = 7$) or BRCA2 ($N_2 = 8$), as described by Hedenfalk *and others* (2001) and analyzed by Efron and Tibshirani (2002) and Efron (2007). The study goal was to identify genes that are differentially expressed in the 2 cancer groups. An important first step toward this goal is to determine whether the data provide evidence against the global null hypothesis that none of the genes are differentially expressed. In this example, each individual's expression levels for different genes are likely to be correlated. The test statistic for the mth gene is the usual 2-sample t-statistic

$$T_m = \frac{\hat{\mu}_{m1} - \hat{\mu}_{m2}}{\left[\frac{1}{7}\sum_{i=1}^{7}(g_{mi} - \hat{\mu}_{m1})^2 + \frac{1}{8}\sum_{i=8}^{15}(g_{mi} - \hat{\mu}_{m2})^2\right]^{1/2}},$$

where g_{mi} is the log-expression level of gene m in the cancer of individual i and $\hat{\mu}_{m1}$ and $\hat{\mu}_{m2}$ are the means in individuals with BRCA1 and BRCA2 mutations, respectively. (The t-distribution assumes that the individual expression levels have lognormal marginal distributions.) We work with the transformed statistics (1.1), where Ψ is the cdf for the t-distribution on $N_1 + N_2 - 2 = 13$ degrees of freedom.

As shown in Table 2, $X = \max_{1 \leqslant m \leqslant 3226} |Z_m| = 4.70$ and $C = 0.4817$. Because this value of C is less than its expected value 0.6827, the histogram of test statistics Z_m is wider than expected under

Table 2. *Effect on overall p-value of correlation among test statistics comparing expression of* 3226 *genes in* 2 *samples of cancer tissue*[†]

Unconditional p-value[‡] P_x, $x = 4.70$	0.006
Conditional p-value[‡] $p_x(c)$, $c = 0.4817$	0.049
var(C)	
Independence	6.71×10^{-5}
Permutation based[‡]	4.04×10^{-3}
Average squared correlation τ^2 [§]	3.36×10^{-2}
cov(C, Y)[¶]	
Independence	-1.76×10^{-6}
Permutation based[‡]	-3.87×10^{-4}
Theoretical[‖]	-9.70×10^{-4}

[†] Seven BRCA1 cancers and 8 BRCA2 cancers from Hedenfalk *and others* (2001).
[‡] Based on $15!/(7!8!) = 6435$ permutations.
[§] From formula (2.3).
[¶] $Y = 1$ if maximum test statistic exceeds 4.70, $Y = 0$ otherwise.
[‖] From formula (2.4).

the assumption of independence. Thus, unconditional overall p-values are likely to be less than conditional ones. To obtain the latter, we considered the $R = 15!/(7!8!) = 6435$ ways of randomly assigning 7 and 8 BRCA1 and BRCA2 labels, respectively, to the 15 patients. We computed the corresponding set of 3226 transformed t-statistics and, for each, its central proportion C_r and a variable Y_r indicating whether its largest test statistic exceeded 4.70, $r = 1, \ldots, 6435$. We then used the regression smoothing of (2.2) to estimate the conditional overall p-value given $C = 0.4817$. The fitted parameters were $\widehat{\alpha} = 3.139$ and $\widehat{\beta} = -12.681$. The unconditional overall p-value, given by the proportion of permutations r with $X_r > 4.70$, was $P_x = P_{4.70} = 0.006$. In contrast, the conditional p-value was $\widehat{p}_x(c) = \widehat{p}_{4.70}(0.4817) = 0.049$. Table 2 gives some insight into the reasons for such a large difference. Shown are the permutation-based value of $\text{var}(C)$ and the estimated mean squared test correlation τ^2 obtained from (2.3). Also shown is the permutation-based value of $\text{cov}(C, Y)$ and its bivariate Gaussian approximation (2.4). The table includes for comparison the values expected if the 3266 test statistics were independent. The correlations among gene expression levels increase $\text{var}(C)$ about 60-fold and increase the magnitude of $\text{cov}(C, Y)$ about 200-fold.

The conditional p-value of 0.049 provides only borderline evidence against the global null hypothesis that none of the genes are differentially expressed, contrary to the interpretation of Hedenfalk *and others* (2001). These conclusions are consistent with those of Efron (2007), who worked with the proportion of false discoveries in these data. Thus, as illustrated by the coin-tossing example of the previous section, conditional and unconditional inferences can be substantially different; when this occurs the conditional ones are more reliable.

4. Discussion

Owen (2005) and Efron (2007) have warned that when the test statistics for a large collection of null hypotheses are correlated, their histogram can be highly variable. This variability can lead to misleading inferences from a single study. This is particularly problematic if the study is large and costly so that replication is difficult. The conditionality principle states that in this situation, inferences conditioned on statistics that capture this variability are more appropriate than unconditional inferences.

Efron (2007) has suggested ways to implement this conditioning when estimating tail probabilities and false discovery rates. The general approach involves defining a statistic X on which inferences are based and then conditioning X on the observed value of a summary statistic C for the histogram spread. Here, we have used this approach to estimate overall significance levels, with X taken as the test statistic with maximum magnitude and C taken as the central proportion of the test statistics. We have focused on overall significance levels to simplify the presentation. However, the proposed permutation-based conditional procedure applies to a large class of inferential problems, such as estimating null tail probabilities, adjusted p-values for subsets of hypotheses of interest, and false discovery rates.

While we have used the central proportion as the conditioning statistic, other choices are possible. However, care is needed in choosing this statistic. In general, the set of M test statistics will be a mixture of 2 unobserved subsets: (i) the test statistics representing the true null hypotheses and (ii) the “nonnull” test statistics representing the false ones. The summary statistic should account for a large fraction of the variability of the histogram of null test statistics and conditioning on it should reduce much of the variability of X. But to avoid power loss by “overconditioning,” the summary statistic should be independent of the nonnull statistics.

Acknowledgments

We are grateful to Joseph B. Keller for helpful comments. *Conflict of Interest:* None declared.

FUNDING

National Institutes of Health (R01 CA94069); National Institute of Mental Health (R01 MH61675, K24 MH64197).

APPENDIX

To prove (2.3) and (2.4), we use the following lemma, which follows from straightforward differentiation.

LEMMA The standard bivariate Gaussian distribution can be expanded in powers of the correlation coefficient ρ:

$$\begin{aligned} f(x, y; \rho) &= \frac{1}{2\pi(1-\rho^2)} \exp\left\{\frac{2\rho - x^2 - y^2}{2(1-\rho^2)}\right\} \\ &= \left[1 + \rho xy + \frac{\rho^2}{2}(1-x^2)(1-y^2)\right]\phi(x)\phi(y) + \mathrm{O}(\rho^3). \end{aligned} \tag{A.1}$$

Here, $\phi(z)$ denotes the standard univariate Gaussian density.

Proof of approximations (2.3) and (2.4). To prove (2.3), we introduce the indicator variable $U_m = 1$ if $|Z_m| < 1$, with $U_m = 0$ otherwise, $m = 1, \ldots, M$. Thus, $E[U_m] = \gamma = \int_{-1}^{1} \phi(z)\,\mathrm{d}z = 0.6827$, and $C = \left(\sum_{m=1}^{M} U_m\right)/M$ has variance

$$\operatorname{var}(C) = \frac{1}{M^2}\sum_{m=1}^{M}\sum_{n=1}^{M} E[U_m U_n] - \gamma^2. \tag{A.2}$$

Our goal is to approximate the expectations $E[U_m U_n]$. We need to only consider pairs $m \neq n$ since $E[U_m^2] = \gamma$. We assume that under H_0, the test statistics Z_m and Z_n have the bivariate Gaussian distribution $f(z_m, z_n; \rho_{mn})$ of (A.1). Then, using the lemma and neglecting terms of order ρ_{mn}^3, we have

$$E[U_m U_n] = \int_{-1}^{1}\int_{-1}^{1} f(x, y; \rho_{mn})\mathrm{d}x\,\mathrm{d}y \sim \gamma^2 + \frac{1}{2}\lambda^2\rho_{mn}^2 = \gamma^2 + 0.11713\rho_{mn}^2, \tag{A.3}$$

where

$$\lambda = \int_{-1}^{1}(1-z^2)\phi(z)\,\mathrm{d}z = 0.4840. \tag{A.4}$$

Thus,

$$E[U_m U_n] \sim \begin{cases} \gamma, & \text{if } m = n, \\ \gamma^2 + 0.11713\rho_{mn}^2, & \text{if } m \neq n. \end{cases} \tag{A.5}$$

Substituting (A.5) into (A.2) gives

$$\begin{aligned} \operatorname{var}(C) &\sim \frac{\gamma(1-\gamma)}{M} + \frac{0.11713}{M^2}\sum_{1 \leqslant m \neq n \leqslant M}\rho_{mn}^2 \\ &= \frac{0.2167}{M} + 0.11713\left(1 - \frac{1}{M}\right)\tau^2 \sim \frac{0.2167}{M} + 0.11713\tau^2, \end{aligned} \tag{A.6}$$

which proves (2.3).

To prove (2.4), we rewrite it as

$$\operatorname{cov}(C, Y) \sim -P_x \left(\frac{\gamma}{M} + \frac{1}{2}\lambda x^2 \tau^2 \right), \tag{A.7}$$

where λ is given by (A.4). Since $C = \left(\sum_{m=1}^{M} U_m\right)/M$, we have

$$\operatorname{cov}(C, Y) = \frac{1}{M} \sum_{m=1}^{M} \operatorname{cov}(U_m, Y) = \frac{1}{M} \sum_{m=1}^{M} E(U_m Y) - \gamma\, P_x. \tag{A.8}$$

But

$$\begin{aligned} E(U_m Y) &= \sum_{n \neq m} \Pr\left(Y = U_m = 1,\, Z_n = \max_m Z_m\right) \\ &= \sum_{n \neq m} \Pr\left(Y = 1,\, Z_n = \max_m Z_m\right) \Pr\left(U_m = 1 \ \mid Y = 1,\, Z_n = \max_m Z_m\right). \end{aligned} \tag{A.9}$$

The first factor in the nth summand of (A.9) is

$$\Pr(Y = 1) \Pr\left(Z_n = \max_m Z_m\right) = P_x/M. \tag{A.10}$$

For large x, the second factor is approximately

$$\Pr\,(U_m = 1 \mid Z_n > x).$$

Substituting these 2 expressions into (A.9) gives

$$E(U_m Y) \sim \frac{P_x}{M} \sum_{n \neq m} \Pr(U_m = 1 \mid Z_n > x) = \frac{P_x}{M} \sum_{n \neq m} A_{mn}, \tag{A.11}$$

where

$$A_{mn} = \frac{1}{\Phi(-x)} \int_{-1}^{1} \int_{x}^{\infty} f(u, v; \rho_{mn}) \mathrm{d}u\, \mathrm{d}v. \tag{A.12}$$

Substituting the expansion (A.1) for the integrand in (A.12) and integrating and ignoring terms of order ρ_{mn}^3 give

$$A_{mn} \sim \gamma - \frac{1}{2}\lambda x^2 \rho_{mn}^2. \tag{A.13}$$

Here, we have used the fact that $\int_x^{\infty} (1 - u^2)\phi(u)\mathrm{d}u = -x\phi(x)$ and that for large values of x, $\Phi(-x)$ can be approximated by $\phi(x)/x$. Substituting (A.13) into (A.11) gives

$$E(U_m Y) = P_x \left(\gamma - \frac{\gamma}{M} - \frac{\lambda x^2}{2M} \sum_{n \neq m} \rho_{mn}^2 \right). \tag{A.14}$$

Finally, substituting (A.14) into (A.8) gives (A.7). □

References

Cox, D. R. and Hinkley, D. V. (1974). *Theoretical Statistics*. London: Chapman and Hall.

Dudbridge, F. and Koeleman, B. (2004). Efficient computation of significance levels for multiple associations in large studies of correlated data, including genomewide association studies. *American Journal of Human Genetics* **75**, 424–435.

Efron, B. (2007). Correlation and large-scale simultaneous significance testing. *Journal of the American Statistical Association* **107**, 93–103.

Efron, B. and Tibshirani, R. (2002). Empirical Bayes methods and false discovery rates for microarrays. *Genetic Epidemiology* **23**, 70–86.

Hedenfalk, I., Duggan, D., Chen, Y., Radmacher, M., Bittner, M., Simon, R., Meltzer, R., Gusterson, B., Esteller, M., Kallioniemi, O. P. *and others* (2001). Gene-expression profiles in hereditary breast cancer. *New England Journal of Medicine* **344**, 539–548.

Owen, A. (2005). Variance of the number of false discoveries. *Journal of the Royal Statistical Society, Series B* **67**, 411–426.

Westfall, P. H. and Young, S. S. (1993). *Resampling-based Multiple Testing*. New York: John Wiley and Sons.

[*Received July 18, 2007; revised November 15, 2007; accepted for publication November 15, 2007*]