Estimating the Number of True Null Hypotheses From a Histogram of p Values

Dan NETTLETON, J. T. Gene HWANG, Rico A. CALDO, and Roger P. WISE

In an earlier article, an intuitively appealing method for estimating the number of true null hypotheses in a multiple test situation was proposed. That article presented an iterative algorithm that relies on a histogram of observed p values to obtain the estimator. We characterize the limit of that iterative algorithm and show that the estimator can be computed directly without iteration. We compare the performance of the histogram-based estimator with other procedures for estimating the number of true null hypotheses from a collection of observed p values and find that the histogram-based estimator performs well in settings similar to those encountered in microarray data analysis. We demonstrate the approach using p values from a large microarray experiment aimed at uncovering molecular mechanisms of barley resistance to a fungal pathogen.

Key Words: False discovery rate; Microarray data; Multiple testing.

1. INTRODUCTION

This article considers the problem of estimating the number of true null hypotheses in a collection of m tests, given the observed p values $p_1, \ldots, p_m$ for the collection of tests. This problem is becoming increasingly important as the analysis of many modern experiments involves testing hundreds, thousands, or even millions of null hypotheses. When mapping quantitative trait loci, for example, each of hundreds of genetic loci are tested for association with a quantitative trait of interest. In microarray experiments, each of thousands of genes are tested for changes in mRNA transcript abundance in response to treatment. Combining the two technologies (see, e.g., Jansen and Nap 2001; Brem, Yvert, Clinton, and Kruglyak 2002; Yvert et al. 2003; Schadt et al. 2003 a,b; Pomp, Allan, and Wesolowski 2004; Hubner et al. 2005; Bystrykh et al. 2005; Chesler et al. 2005; DeCook, Lall, Nettleton, and Howell 2006) can result in millions of tests. In such situations it is natural to attempt to estimate

Dan Nettleton is Associate Professor, Department of Statistics, Iowa State University, Ames, IA 50011-1210 (E-mail: *dnett@iastate.edu*). J. T. Gene Hwang is Professor, Departments of Mathematics and Statistics, Cornell University, Ithaca, NY 14853-4201. Rico A. Caldo is Postdoctoral Research Associate, Department of Plant Pathology and Center for Plant Responses to Environmental Stresses, Iowa State University, Ames, IA 50011-1020. Roger P. Wise is Professor, Department of Plant Pathology and Center for Plant Responses to Environmental Stresses, and Research Plant Geneticist, USDA-ARS-Corn Insects and Crop Genetics Research Unit, Iowa State University, Ames, IA 50011-1020.

Journal of Agricultural, Biological, and Environmental Statistics, Volume 11, Number 3, Pages 337–356
DOI: 10.1198/108571106X129135

the number of true or, equivalently, the number false null hypotheses.

As a specific example, consider the microarray experiment conducted by Caldo, Nettleton, and Wise (2004) to examine gene expression in barley plants during fungal infection. Two fungal isolates were used to infect plants of three barley genotypes. Affymetrix GeneChips (Close et al. 2004; Lipshutz, Fodor, Gingeras, and Lockhart 1999) were used to measure gene expression in plants at six time points following fungal inoculation. Full details of the experimental design were provided by Caldo et al. (2004).

The main purpose of the experiment can be described as follows. Prior to the experiment it was known that some combinations of fungal isolate and barley genotype result in what is known as a compatible interaction in which the barley plants are susceptible to colonization by the fungal isolate, resulting in widespread damage to plant cells. Other isolate-genotype combinations were known to result in incompatible interactions in which barley plants are resistant to infection by the fungal isolate, resulting in virtually no damage to plant cells. Although the phenotypes of certain barley genotype-fungal isolate combinations are known, the molecular mechanisms underlying incompatible and compatible interactions are not completely understood. Because of the complexity of the plant-fungus interactions, it was assumed that the genetic pathways involved in these biological phenomena consisted of multiple genes that worked to make plants resistant or susceptible to fungal attack. To learn about the genes involved in these genetic pathways, Caldo et al. sought genes whose mean expression pattern over the course of infection differed in compatible and incompatible interactions.

Contrasts for nonparallel expression patterns in compatible and incompatible interactions were constructed as part of a mixed linear model analysis conducted separately for each of 22,840 genes. Full details of the data analysis were provided by Caldo et al. (2004). Figure 1 shows two portions of results from the analysis that are relevant for our current article. Figure 1(a) depicts the gene with the smallest p value for the contrast of interest, and Figure 1(b) shows the histogram of p values from all 22,840 contrasts. The gene in Figure 1(a) exhibits similar expression in compatible and incompatible interactions up through 16 hours after inoculation. Between 16 and 20 hours mean expression seems to drop in compatible interactions relative to the mean level maintained in incompatible interactions. Because of this nonparallel expression pattern, we have good reason to suspect that the behavior of this gene plays a role in determining whether plants will be resistant or susceptible to the fungus. How many other genes have expression patterns that might suggest a similar role? That is the question we wish to address in our current work. We will present a method for using the information in the histogram of Figure 1(b) to estimate the number of true and false null hypotheses among the 22,840 null hypotheses tested.

This article focuses on a histogram-based method for estimating the number of true null hypotheses originally proposed by Mosig et al. (2001). The basic method is described in Section 2. Informal and formal descriptions of an iterative algorithm proposed by Mosig et al. (2001) are contained in Sections 2.1 and 2.2, respectively. Section 2.3 contains a simple illustrative example. Section 2.4 shows that the iterative algorithm converges to a limit as the number of iterations approaches infinity. Furthermore, we characterize the limit

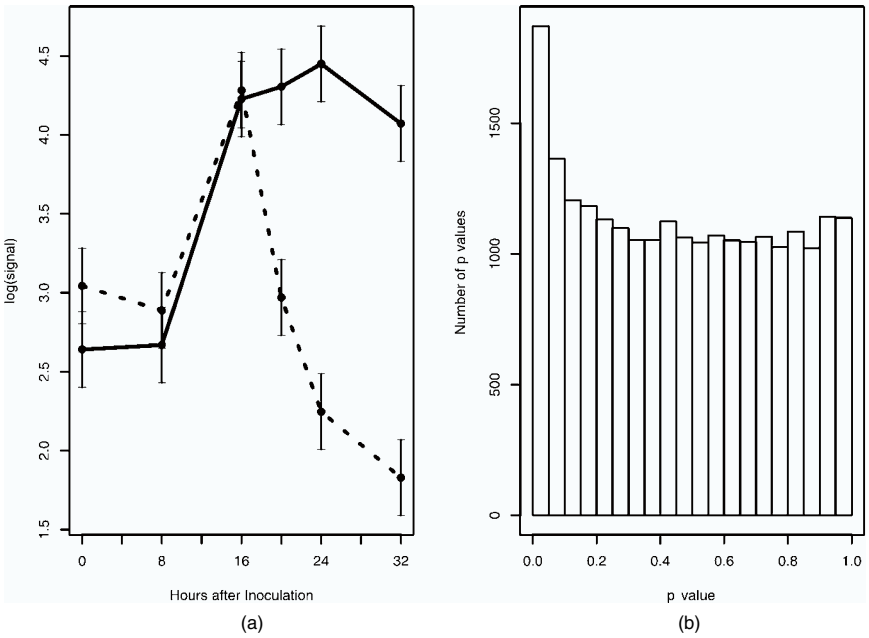

Figure 1. (a) Estimated mean expression plus or minus one standard error for a gene during incompatible (solid line) and compatible (dashed line) barley-fungus interactions. (b) Histogram of 22,840 p values for comparing compatible and incompatible expression patterns in the barley experiment.

and show that the estimator can be computed directly from the observed p values with no iteration. Section 2.5 demonstrates the approach using the p values depicted in Figure 1(b) from the barley experiment. Section 3 describes other approaches to the problem. Section 4 presents a method for automatically choosing the number of bins used by the histogram-based estimator. Section 5 describes a simulation study that compares the estimator with other approaches for estimating the number of true null hypotheses from the distribution of observed p values. Section 6 provides conclusions and ideas for future work. The article concludes with an Appendix containing all proofs.

2. AN INTUITIVE ALGORITHM FOR ESTIMATING THE NUMBER OF TRUE NULL HYPOTHESES

Suppose a collection of m null hypotheses $H_{01}, \ldots, H_{0m}$ is to be tested with continuously distributed test statistics $t_1, \ldots, t_m$. Suppose that m_0 of the null hypotheses are true, and denote the proportion of true null hypotheses (m_0/m) by π_0. We assume that, for all $\alpha \in (0, 1)$, each null hypothesis is tested using a test for which the probability of rejecting a true null hypothesis is α whenever the test is conducted at significance level α. Then the p values corresponding to tests with true null hypotheses will follow a uniform distribution on the interval $(0, 1)$. Further suppose that the tests used are unbiased, implying that the probability of rejecting a false null hypothesis with a significance-level-α test will

be greater than or equal to α. (Note that the rejection probability must be greater than α in some regions of the alternative space so that the rejection probability is not simply equal to a constant α regardless of the data.) When these conditions are satisfied, the distribution of observed p values from the m tests will be a mixture of a uniform distribution and a distribution that is stochastically smaller than uniform. The mixing proportion associated with the uniform distribution is π_0.

Mosig et al. (2001) proposed an iterative algorithm that essentially estimates the proportion of observed p values that follow a uniform distribution. By examining the examples presented by Mosig et al. (2001), we were able to arrive at the following description of their algorithm that can be used to reproduce the results in Table 2 of Mosig et al. (2001).

2.1 INFORMAL DESCRIPTION OF THE ITERATIVE ALGORITHM

1. The interval $[0, 1]$ is partitioned into B bins of equal width.
2. Initially assume that all null hypotheses are true ($m_0 = m$).
3. Find the expected number of p values for each bin given the current estimate of the number of m_0, bearing in mind that p values corresponding to true null hypotheses should be uniformly distributed on the interval $(0, 1)$.
4. Beginning with the leftmost bin, sum the number of p values in excess of expected until a bin with no excess is reached.
5. The excess sum is an estimate of the number of false null hypotheses. Subtracting the excess sum from m provides a new estimate of m_0.
6. Return to Step 3 and repeat until convergence.

2.2 FORMAL DEFINITION OF THE ITERATIVE ALGORITHM

We now give a formal definition of this algorithm with notation that will be convenient for stating our main results and proving those results in the Appendix. A simple example that should help to clarify the notation immediately follows the formal definition.

Suppose the interval $[0, 1]$ is partitioned into B bins numbered from 1 to B such that bin 1 is $[0, 1/B]$ and bin i is $(\frac{i-1}{B}, \frac{i}{B}]$ for $i = 2, \ldots, B$. Let n_i denote the number of p values falling into bin i for $i = 1, \ldots, B$. Let $\bar{n}_{i:B} = \sum_{j=i}^{B} n_j/(B - i + 1)$ for any $i = 1, \ldots, B$. Let $N_0 = \sum_{i=1}^{B} n_i = B\bar{n}_{1:B}$ and define

$$\frac{N_k}{B} = \left(\frac{i_k - 1}{B}\right) \frac{N_{k-1}}{B} + \left(1 - \frac{i_k - 1}{B}\right) \bar{n}_{i_k:B} \quad \text{for all} \quad k \geq 1, \tag{2.1}$$

where

$$i_k \equiv \min \left\{ i : n_i \leq \frac{N_{k-1}}{B} \right\}.$$

Then N_k is the estimated number of true nulls at iteration k of the algorithm introduced by Mosig et al. (2001).

2.3 A SIMPLE EXAMPLE

Suppose 100 tests are conducted with observed p values distributed as follows: 36 p values in the interval $[0.0, 0.2]$, 22 p values in $(0.2, 0.4]$, 20 p values in $(0.4, 0.6]$, 10 p values in $(0.6, 0.8]$, and 12 p values in $(0.8, 1.0]$. In this example $B = 5$, $n_1 = 36$, $n_2 = 22$, $n_3 = 20$, $n_4 = 10$, $n_5 = 12$, and $N_0 = 100$. In terms of the informal description of the algorithm, we see that in the first bin there is and excess of $36 - 20 = 16$ p values over the expected number of p values under the assumption that $N_0 = 100$ tests have true null hypotheses. Likewise there is an excess of $22 - 20 = 2$ p values in the second bin. Because there is no excess in the third bin, our updated estimate of m_0 is $N_1 = 100 - (16 + 2) = 82$. Now we should expect $82/5 = 16.4$ p values per bin to correspond to tests with true null hypotheses. We have excesses in the first three bins of 19.6, 5.6, and 3.6, respectively. Thus, $N_2 = 100 - (19.6 + 5.6 + 3.6) = 71.2$. Continuing this process, it is straightforward to show that N_k grows arbitrarily close to 55.

In terms of the formal algorithm, we have $i_1 = 3$ because $n_1 = 36$ and $n_2 = 22$ are both greater than $N_0/B = 20$ and $n_3 = 20 \leq 20 = N_0/B$. Thus, by (2.1) we have

$$N_1 = 5\left\{\left(\frac{2}{5}\right)\frac{100}{5} + \left(1 - \frac{2}{5}\right)\frac{20 + 10 + 12}{3}\right\} = 82.$$

Similarly, it is straightforward to show that $i_k = 4$ for all $k > 1$ and $N_2 = 71.2$, $N_3 = 64.72$, $N_4 = 60.832$, $N_5 = 58.499$, $N_6 = 57.100$, $N_7 = 56.260$, $N_8 = 55.756$, $N_9 = 55.453$, $N_{10} = 55.272$, $N_{11} = 55.163$, $N_{12} = 55.098$, $N_{13} = 55.059$, etc. Due to round-off error the algorithm will converge in a finite number of steps to an estimate of 55 for the number of true null hypotheses. The first few steps of the iterative procedure and its limit are depicted in Figure 2. The horizontal dashed lines corresponding to the values of N_k/B can be viewed as iterative estimates of the uniform distribution for the p values corresponding to tests with true null hypotheses. At each step of the algorithm, the discrepancy between the observed distribution and the proposed uniform component of the observed distribution is used to obtain a new estimate of the uniform component.

2.4 EXISTENCE AND CHARACTERIZATION OF THE LIMIT

The iterative algorithm described in the previous subsection appears to converge quickly to an estimate in a finite number of steps in all the examples that we have investigated. The following result states that a limit does exist and characterizes that limit as a simple function of the bin counts associated with the distribution of observed p values. The proof (contained in the Appendix) shows that the limit is usually not achieved in a finite number of iterations, and the apparent convergence is due to round-off error (as was the case in the example of the previous section).

Convergence Result: Let $I = min\{i : n_i \leq \bar{n}_{i:B}\}$. *Then* $lim_{k\to\infty} N_k = B\bar{n}_{I:B}$.

Note that in the example of the previous subsection $n_1 = 36 > 20 = \bar{n}_{1:5}$, $n_2 = 22 > 16 = \bar{n}_{2:5}$, $n_3 = 20 > 14 = \bar{n}_{3:5}$, and $n_4 = 10 \leq 11 = \bar{n}_{4:5}$. Thus, $I = 4$, and the convergence result implies that the limiting value for the iterative algorithm is

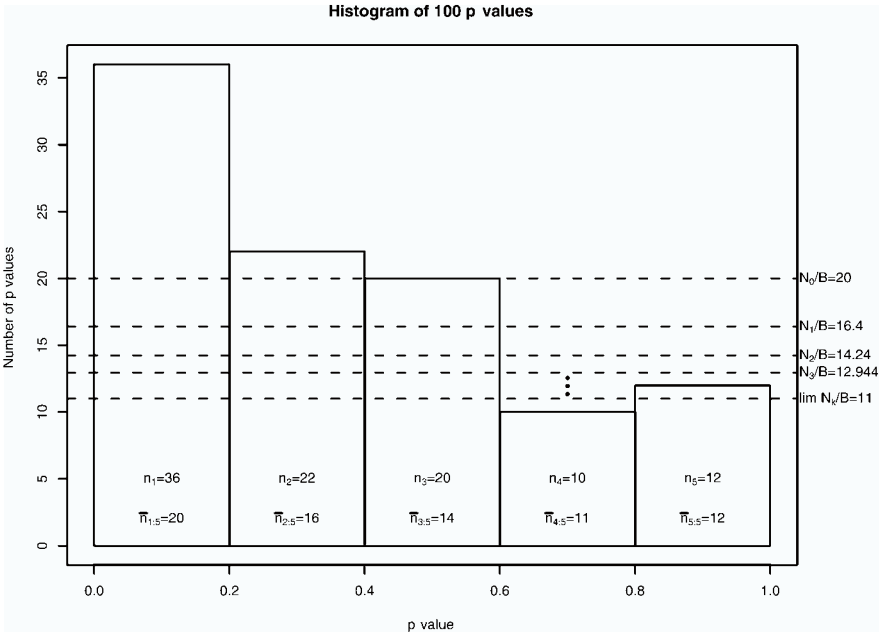

Figure 2. Histogram of 100 p values for a simple example illustrating the proposed method for estimating the number of true null hypotheses. The top four dashed lines represent the first four iterative estimates of the expected number of p values per bin believed to correspond to true null hypotheses. The lowest dashed line represents the limit of the iterative procedure. The labels on the bars of the histogram correspond to the observed number of p values per bin and the average of bin counts over the current bin and all bins to the right. The convergence theorem of Section 3 dictates that $\lim_{k\to\infty} N_k/B = \bar{n}_{4:5} = 11$ *because the fourth bin is the first satisfying* $n_i \leq \bar{n}_{i:B}$.

$B\bar{n}_{I:B} = 5\bar{n}_{4:5} = 5(11) = 55$, as was observed in the example.

Though neither approach is computationally complex, using the convergence result to compute an estimate of m_0 clearly requires far fewer steps than the iterative algorithm proposed by Mosig et al. (2001). R code for computing the estimate of m_0 from a vector of observed p values is available on the first author's Web site at *http://www.public.iastate.edu/~dnett/m0estimation.shtml*. Users may specify the number of bins used to obtain the estimate (the default is $B = 20$). Alternatively, a bootstrap approach described in Section 4 may be used to automatically select the number of bins. The R code will produce an estimate for an observed vector of over 20,000 p values using $B = 20$ bins in less than a second on a typical personal computer. The bootstrap procedure is far more computationally intensive with an execution time that depends on the extent of bootstrap resampling and the candidate bin choices specified by the user.

2.5 APPLICATION TO THE BARLEY DATA

Figure 3(a) shows the histogram of the 22,840 p values from the analysis of the barley data along with a dashed line indicating the estimate of the uniform portion of the distribution that is obtained using our histogram-based estimator with $B = 20$ bins. The seventh bin is

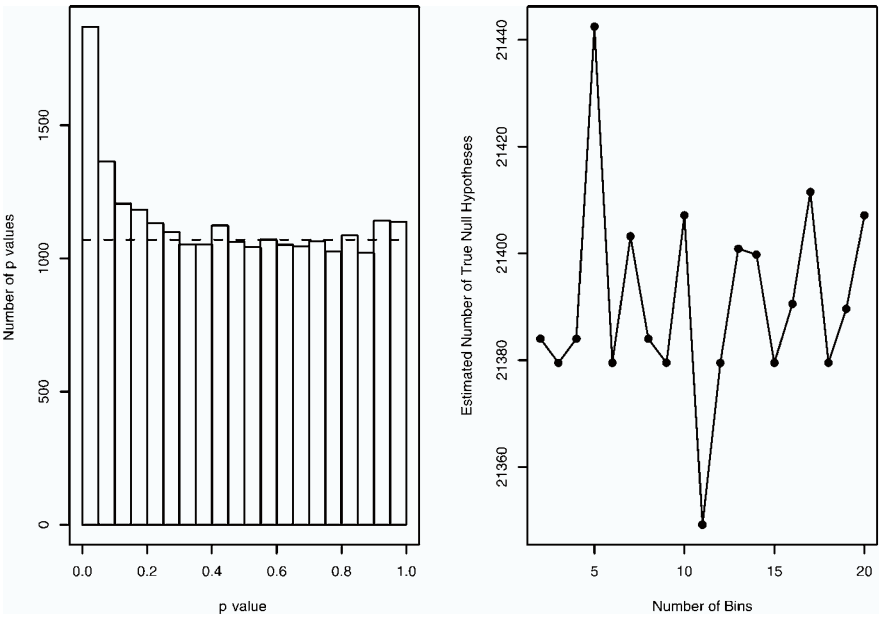

Figure 3. (a) Histogram of 22,840 p values for comparing expression patterns in the barley experiment. The dashed line represents the uniform distribution corresponding to p values for nondifferentially expressed genes as estimated by the noniterative method described in Section 2. (b) Histogram-based estimates of the number of true null hypotheses as a function of the number of histogram bins.

the leftmost of the 20 bins with a count less than or equal to its tail average ($n_7 = 1053 \leq 1070.4 = \bar{n}_{7:20}$); thus, m_0 is estimated to be $20 * 1070.4 = 21{,}408$ by the 20-bin estimator.

Figure 3(b) shows how the histogram-based estimates of m_0 change for $B = 2, 3, \ldots,$ 20 bins. Although at first glance it may appear that the estimator of m_0 is quite sensitive to the number of bins, note that these estimates range from 21,349 to 21,443, and the estimates of π_0 fall within the very narrow range of 0.935 to 0.939. Although bin choice is not a major issue in this case, we can use the bootstrap approach to be described in Section 4 to automatically choose the number of bins. The algorithm selected $B = 20$ (our default choice) for the barley data based on 1,000 with-replacement samples of size 22,840 from the observed p values.

The results indicate that somewhere in the neighborhood 1,432 genes may have expression patterns that differ between compatible and incompatible interactions. Caldo et al. (2004) identified 22 genes whose expression patterns and p values (≤ 0.0001) suggested a role in distinguishing compatible and incompatible interactions. Clearly the results presented here suggest that there are many other genes of interest. However, note that although we have estimated that approximately 1,432 genes may be of interest, we cannot know based solely on their p values which 1,432 of 22,840 genes are the genes of interest. With approximately 21,408 genes with true null hypotheses, we should expect many very small p values to correspond to genes with true null hypotheses. Also many of the differentially

expressed genes are likely to have relatively large p values if their level of differential expression is small.

The inevitable mixing of p values for genes with true and false null hypotheses can be seen in Figure 3(a), where the portion of the distribution stochastically smaller than uniform appears to extend up to p values as large as 0.3 although the vast majority of genes with p values under 0.3 appear to come from the uniform component of the distribution. Identifying more of the 1,432 genes of interest requires either a larger experiment, more powerful analysis strategies, or the use of additional biological information regarding gene function. The ability to estimate the number differentially expressed genes is important here or in any microarray analysis because it helps researchers understand how many additional genes might be involved in the process of interest and suggests how much additional discovery might be possible with follow-up research.

3. OTHER APPROACHES FOR ESTIMATING THE NUMBER OF TRUE NULL HYPOTHESES

Several approaches have been proposed for estimating the number of true null hypotheses when conducting many tests. Much of the recent work in this area has been motivated by the attempts to control or estimate error rates related to the proportion of mistakenly rejected null hypotheses among all rejected null hypotheses. Benjamini and Hochberg (1995) presented the seminal work in this area. They showed that a sequential testing procedure proposed by Simes (1986) provides control of the false discovery rate (FDR), formally defined as $E(Q)$ where Q is the proportion of mistakenly rejected null hypotheses among all rejected null hypotheses or 0 if no hypotheses are rejected. As noted by Benjamini and Hochberg (1995), the sequential procedure of Simes (1986) actually provides control at π_0 times the nominal rate. Thus, it is possible to improve procedures aimed at FDR control by using information about π_0 or, equivalently, m_0. Benjamini and Hochberg (2000), Mosig et al. (2001), Storey (2002a, b), Storey and Tibshirani (2003), Fernando et al. (2004), and Genovese and Wasserman (2004) among others have presented methods that use estimates of m_0 to estimate or control quantities related to FDR.

This section describes several methods for estimating the number of true null hypotheses. We will compare these approaches to the histogram-based estimator in a simulation study described in Section 4. It should be noted that Hsueh, Chen, and Kodell (2003) and Nguyen (2004) have conducted simulation studies to assess the performance of several methods for estimating m_0; however, neither article considers the histogram-based estimator considered here nor many of the more recent approaches described in the following.

3.1 THE LOWEST SLOPE ESTIMATOR

More than 20 years ago, Schweder and Spjøtvoll (1982) proposed a graphical method for estimating the number of true null hypotheses from a display of the observed p values equivalent to a plot of p values against their ranks. Hochberg and Benjamini (1990) used

ideas presented by Schweder and Spjøtvoll to develop the lowest slope estimator of the number of true null hypotheses. A detailed description and motivation for this estimator was also provided by Benjamini and Hochberg (2000) and Hsueh et al. (2003). The estimator is given by $\min\{1/S_J + 1, n\}$, where $S_j = \frac{1-p_{(j)}}{n+1-j}$, $p_{(j)}$ denotes the jth smallest p value, and J is the smallest integer for which $S_J < S_{J-1}$. The name comes from the fact that S_j is the slope of the line passing through the points $(n+1, 1)$ and $(j, p_{(j)})$, and S_J is the first slope that is lower than it predecessor when working from the smallest p value to the largest.

3.2 A CONVEX DENSITY ESTIMATOR

Recently Langaas, Ferkingstad, and Lindqvist (2005) developed and compared several methods for estimating the proportion of true null hypotheses based on estimates of the density of the observed p value distribution. Their preferred estimator is based on the nonparametric maximum likelihood estimate of the p value density subject to the restriction that the estimated density be decreasing and convex. The height of the density at 1 is used to estimate m_0.

3.3 THE λ-ESTIMATORS

Storey (2002a) proposed to estimate the number of true null hypotheses using an estimator equivalent to

$$\hat{m}_0(\lambda) = \frac{\sum_{j=1}^{m} x_j(\lambda)}{1-\lambda}, \tag{3.1}$$

where $x_j(\lambda) = 1$ if $p_j > \lambda$ and 0 otherwise. Motivation for this estimator was originally discussed by Schweder and Spjøtvoll (1982). As a simple example, consider $\hat{m}_0(0.75)$. If we assume for the moment that only p values from tests with true null hypotheses will exceed 0.75, the numerator in (3.1) is a count of the number of p values from tests with true null hypotheses that fall in the last fourth of the interval $[0, 1]$. Because p values from tests with true null hypotheses are assumed to be uniformly distributed, we should multiply this count by 4 (equivalently divide by $1 - 0.75$) to estimate the total number of tests with true null hypotheses.

The estimator $\hat{m}_0(\lambda)$ will tend to overestimate m_0 because p values from tests with false null hypotheses will exceed λ with positive probability. The positive bias in the estimator can be reduced by choosing values of λ near 1, but this will clearly increase the variation in the estimator due to division by a smaller value of $1 - \lambda$. To balance the bias-versus-variance trade off, Storey (2002a) proposed a bootstrap approach for choosing a value of λ to minimize the mean squared error of the $\hat{m}_0(\lambda)$. The idea behind the approach is to approximate

$$\text{MSE}(\lambda) = E\{\hat{m}_0(\lambda) - m_0\}^2 \quad \text{by} \quad \widehat{\text{MSE}}(\lambda) = \frac{1}{K}\sum_{k=1}^{K}\{\hat{m}_0^{*k}(\lambda) - \min_{\lambda}\hat{m}_0(\lambda)\}^2,$$

where $\hat{m}_0^{*k}(\lambda)$ represents the kth of K bootstrap replications of $\hat{m}_0(\lambda)$ obtained by drawing a with-replacement simple random sample of size m from the observed p values, and $\min_\lambda \hat{m}_0(\lambda)$ denotes the minimum of $\hat{m}_0(\lambda)$ over $\lambda = 0, 0.05, \ldots, 0.95$ or a similar sequence of values over $[0, 1)$. The function $\widehat{\text{MSE}}(\lambda)$ is calculated for $\lambda = 0, 0.05, \ldots, 0.95$ or a similar sequence of values over $[0, 1)$. The value of λ that minimizes $\widehat{\text{MSE}}(\lambda)$, say $\hat{\lambda}$, is used to produce $\hat{m}_0(\hat{\lambda})$ as an estimate of m_0.

Rather than choosing a single value of λ, Storey and Tibshirani (2003) proposed an estimator derived from the fit of a natural cubic spline with three degrees of freedom to points $(\lambda, \hat{m}_0(\lambda))$ for a sequence of λ values over $[0, 1)$. In particular they estimated m_0 by the value of the fitted natural cubic spline at $\lambda = 1$.

4. CHOOSING THE NUMBER OF BINS FOR THE HISTOGRAM-BASED ESTIMATOR

We begin this section by showing that the histogram-based estimator can be written as an estimator of the form $\hat{m}_0(\lambda)$ described in Section 3.3. Using notation from Sections 2.4 and 3.3, we have

$$B\bar{n}_{I:B} \quad = \quad B \cdot \frac{n_I + \cdots + n_B}{B - I + 1} = \frac{\sum_{j=1}^{m} x_j \left(\frac{I-1}{B}\right)}{\frac{B-I+1}{B}} = \frac{\sum_{j=1}^{m} x_j \left(\frac{I-1}{B}\right)}{1 - \frac{I-1}{B}} = \hat{m}_0\left(\frac{I-1}{B}\right).$$

Thus, it follows from the convergence result in Section 2.4 that the histogram-based estimator is a λ-estimator with $\lambda = \frac{I-1}{B}$.

Recall that I is a function of the observed p values. Thus the histogram-based estimator offers an approach for choosing a value of λ as a function of the observed data. Specifying a number of bins is equivalent to providing a list of candidate λ values from which the histogram-based algorithm is allowed to choose. For example, choosing $B = 20$ restricts λ to be selected from the set $\{0, 0.05, \ldots, 0.95\}$.

In Section 3.3, we discussed how the choice of λ affects the variance and bias of the estimator $\hat{m}_0(\lambda)$. Similarly, the choice of B has an impact on the variance and bias of the histogram-based estimator. We can attempt to manage the variance-versus-bias trade off in our selection of B using an approach completely analogous to that used by Storey (2002a) for selection of λ. The approach is described as follows:

Let $\tilde{m}_0(B)$ denote the histogram-based estimate of m_0 obtained when using B bins, and let $\tilde{m}_0^{*k}(B)$ represent the kth of K bootstrap replications of $\tilde{m}_0(B)$ obtained by drawing a with-replacement simple random sample of size m from the observed p values. Compute

$$\widetilde{\text{MSE}}(B) = \frac{1}{K} \sum_{k=1}^{K} \{\tilde{m}_0^{*k}(B) - \min_B \tilde{m}_0(B)\}^2,$$

where $\min_B \tilde{m}_0(B)$ denotes the minimum of $\tilde{m}_0(B)$ over $B = 2, 3, \ldots, 20$ or a similar sequence of choices for B. The value of B in this same range that minimizes $\widetilde{\text{MSE}}(B)$ is selected as the number of bins used to obtain the final estimate of m_0. We explore the performance of this approach via simulation in the next section.

5. A SIMULATION STUDY AND PERFORMANCE COMPARISON

5.1 SIMULATION STRUCTURE

The mean squared error (MSE) and bias of seven estimators of the number of true null hypotheses were examined through a simulation study. The factors considered in the study were the proportion of true null hypotheses (50, 75, 90, and 100%), power to detect departures from the null hypothesis, and the correlation structure among the tests. The simulation was designed to be similar to situations encountered in the analysis of microarray data. Each simulated dataset consisted of expression levels of 10,000 genes for each of three control and three treated experimental units. The six 10,000-dimensional data vectors were distributed as independent multivariate normal vectors with a common covariance structure and a mean that was allowed to depend on the control or treatment status of each experimental unit.

Data for all 10,000 genes in any one experimental unit were generated as either mutually independent (IND) or with a block correlation structure and blocks of size 100. We considered two different block correlation structures: compound symmetry (CSY) with pairwise correlation of 0.9 between any pair of elements within a block and an autoregressive order one structure (AR1) with correlation of $(-0.9)^{|i-j|}$ for the ith and jth observations within any block. Block structures have been used to simulate correlated microarray data in past work (e.g., see Storey 2002b) because genes are believed to work together in functional groups known as pathways. The expression of genes working together in a pathway are expected to be correlated while genes in different pathways may work independently of one another. The CSY structure was chosen to judge the impact of extreme positive correlation among genes even though such structures are not expected in practice. The AR1 structure was chosen to mimic more biologically relevant situations where a mix of positive and negative correlations of varying magnitude is expected.

The p values used to estimate the number of null hypotheses for each simulated dataset were generated by conducting a two-sample t test for each gene. The treatment effects necessary for an individual 0.05-level t test to have powers of 0.5, 0.7, or 0.9 (say $\delta_{0.5}$, $\delta_{0.7}$, and $\delta_{0.9}$, respectively) were determined and used to set the means of the multivariate normal distributions used to simulate data for cases where the proportion of true null hypotheses was less than one. More specifically we considered three power scenarios. In the "small effects" case, treatment effects for differentially expressed genes (i.e., genes whose treatment and control means differed) were simulated by drawing independent gamma random variables with variance 1 and mean $\delta_{0.5}$. The "medium effects" and "large effects" scenarios were constructed in the same manner except that $\delta_{0.5}$ was replaced by $\delta_{0.7}$ or $\delta_{0.9}$, respectively. Allowing for variation among the treatment effects in any one dataset provides a better match to the situations encountered in practice where no two treatment effects can be expected to be exactly the same.

Five hundred datasets were generated for each of 30 conditions. Twenty-seven conditions were defined by all possible combinations of number of true null hypotheses (50, 75, or

90% of 10,000 tests), correlation structure (IND, CSY, or AR1), and power scenario (small, medium, or large effects). The other three scenarios were obtained by considering the three correlation structures in the case where all 10,000 null hypotheses were true. The number of true null hypotheses was estimated for the 30×500 datasets using the following methods: the lowest slope estimator described in Section 3.1 (LS), the convex density estimator described in Section 3.2 (CD), the cubic spline approach described in Section 3.3 (CS), the λ-estimator with the value of λ selected via bootstrapping as described in Section 3.3 (LB), the histogram-based estimator with $B = 20$ bins (H20), and the histogram-based estimator with B selected from $2, 3, \ldots, 20$ via the bootstrap approach described in Section 4 (HB). Empirical estimates of mean squared error and bias were obtained for each combination of method and condition.

5.2 SIMULATION RESULTS AND DISCUSSION

Table 1 shows the square root of the estimated mean squared error (RMSE) across all 30 simulation conditions for the six methods considered in the simulation study. The lowest three estimated RMSEs for each condition are printed in bold. The lowest RMSE for each condition is highlighted with a box while the second lowest is underlined. The 6 methods were ranked for each condition, and the average rank across all 30 conditions was determined for each method as follows: CD = 1.93, H20 = 2.77, HB = 2.80, LB = 3.40, CS= 4.90, and LS = 5.20.

Table 2 shows the estimated bias across all 30 simulation conditions for the six methods studied. Negative bias estimates are printed in bold. Note that negative bias is arguably more problematic than positive bias because a negative bias could lead to under estimation of FDR or other related error measures designed to control Type I error.

The 500 replications of each simulation condition can be used to evaluate whether the differences observed between the methods were significant beyond Monte Carlo error. In the large majority of the comparisons (greater than 94%), a method as simple as the Wilcoxon signed rank test on the errors or squared errors indicates significant differences between a pair of methods with respect to estimated bias or MSE values at the 0.05 level.

The LS methods tended to exhibit the highest mean squared error among all the methods studied. Its RMSE estimates were often more than two times larger than those of the top performing methods and were more than 10 times larger for some of the simulation settings. The method tended to substantially overestimate m_0 except, of course, when all null hypotheses were true. When $m = m_0$, the LS method exhibited very little bias and very low MSE compared to the other five methods.

The performance of the LS estimator is not surprising given that Benjamini and Hochberg did not develop the estimator with the intent of achieving low MSE for cases when $\pi_0 < 1$. Rather their goal was to develop a conservative estimator of π_0 so that their adaptive method for controlling FDR would maintain control across a wide variety situations, including situations when $\pi_0 \approx 1$.

The CS method typically had lower RMSE than the LS method, but it was dominated in

Table 1. Estimated Root Mean Squared Error for Estimators of m_0. CD = convex density, H20 = histogram 20 bins, HB = histogram bootstrap, LB = λ-bootstrap, CS = cubic spline, LS = lowest slope.

Correlation structure	*Proportion of true nulls*	*Effect sizes*	*Estimated root mean squared error*					
			CD	*H20*	*HB*	*LB*	*CS*	*LS*
IND	1.00	0	**81**	**34**	94	267	208	**1**
IND	0.90	small	**174**	**230**	**178**	250	325	942
IND	0.90	medium	**114**	**146**	**136**	285	315	848
IND	0.90	large	**105**	**70**	**126**	266	307	614
IND	0.75	small	**432**	504	**421**	**394**	491	2122
IND	0.75	medium	**168**	223	**178**	**218**	280	1683
IND	0.75	large	**97**	**90**	**115**	237	265	1022
IND	0.50	small	787	842	**761**	**717**	**725**	3577
IND	0.50	medium	**278**	344	**271**	**265**	314	2528
IND	0.50	large	**85**	**97**	**103**	183	213	1290
AR1	1.00	0	**219**	**154**	245	380	334	**1**
AR1	0.90	small	**274**	**293**	**290**	334	439	941
AR1	0.90	medium	**243**	**242**	**271**	360	448	847
AR1	0.90	large	**224**	**182**	**261**	343	415	625
AR1	0.75	small	**445**	498	**440**	**417**	515	2122
AR1	0.75	medium	**251**	**291**	**268**	299	411	1694
AR1	0.75	large	**216**	**187**	**246**	335	389	1036
AR1	0.50	small	827	874	**806**	**766**	**793**	3619
AR1	0.50	medium	**339**	385	**339**	**333**	398	2576
AR1	0.50	large	**184**	**169**	**202**	269	311	1286
CSY	1.00	0	**730**	**579**	794	986	1012	**12**
CSY	0.90	small	**723**	**784**	**819**	888	1046	936
CSY	0.90	medium	**749**	**762**	**829**	895	1055	851
CSY	0.90	large	**721**	**748**	862	971	1124	**644**
CSY	0.75	small	**774**	862	**815**	**816**	1223	2113
CSY	0.75	medium	**684**	**749**	**753**	784	1145	1700
CSY	0.75	large	**657**	**703**	**749**	825	1155	1044
CSY	0.50	small	**989**	1043	**976**	**946**	1215	3650
CSY	0.50	medium	**660**	720	**680**	**669**	964	2579
CSY	0.50	large	**566**	**606**	**626**	664	929	1288

Table 2. Estimated Bias for Estimators of m_0. CD = convex density, H20 = histogram 20 bins, HB = histogram bootstrap, LB = λ-bootstrap, CS = cubic spline, LS= lowest slope.

Correlation	*Proportion of*	*Effect*	*Estimated root mean squared error*					
structure	*true nulls*	*sizes*	*CD*	*H20*	*HB*	*LB*	*CS*	*LS*
IND	1.00	0	**−52**	**−19**	**−56**	**−159**	**−118**	**−1**
IND	0.90	small	126	209	113	**−8**	128	942
IND	0.90	medium	18	87	9	**−106**	28	846
IND	0.90	large	**−46**	10	**−50**	**−144**	10	609
IND	0.75	small	411	480	389	321	399	2120
IND	0.75	medium	117	183	100	9	93	1678
IND	0.75	large	**−34**	33	**−26**	**−113**	14	1015
IND	0.50	small	776	823	743	683	684	3570
IND	0.50	medium	252	314	234	183	214	2516
IND	0.50	large	3	62	**−2**	**−71**	20	1276
AR1	1.00	0	**−127**	**−67**	**−137**	**−249**	**−197**	**−1**
AR1	0.90	small	86	170	71	**−18**	105	940
AR1	0.90	medium	**−22**	44	**−44**	**−130**	31	844
AR1	0.90	large	**−96**	**−33**	**−115**	**−186**	**−8**	620
AR1	0.75	small	367	427	343	274	352	2120
AR1	0.75	medium	98	156	75	4	116	1688
AR1	0.75	large	**−61**	5	**−64**	**−145**	**−8**	1027
AR1	0.50	small	788	834	761	708	713	3610
AR1	0.50	medium	250	313	232	169	203	2563
AR1	0.50	large	**−27**	34	**−31**	**−100**	14	1273
CS	1.00	0	**−429**	**−245**	**−469**	**−657**	**−605**	**−2**
CS	0.90	small	**−120**	**−57**	**−203**	**−313**	**−73**	932
CS	0.90	medium	**−288**	**−203**	**−374**	**−460**	**−199**	840
CS	0.90	large	**−343**	**−246**	**−455**	**−559**	**−239**	624
CS	0.75	small	193	218	128	58	333	2102
CS	0.75	medium	**−95**	**−61**	**−152**	**−216**	111	1677
CS	0.75	large	**−288**	**−242**	**−353**	**−437**	**−19**	1017
CS	0.50	small	710	718	659	602	725	3631
CS	0.50	medium	132	170	91	46	194	2550
CS	0.50	large	**−197**	**−152**	**−230**	**−301**	**−14**	1253

our simulation by the LB approach. Recall that both the CS and LB methods are determined from λ-estimators as described in Section 3.3. The CS approach uses a natural cubic spline to combine multiple λ-estimators into a single estimator while the LB approach uses a bootstrapping procedure to identify a good choice for λ. The LB approach often exhibited the lowest RMSE among all methods when many genes were differentially expressed with small effect sizes, but it also had the greatest degree of negative bias among all methods considered. This negative bias is perhaps not surprising given that $\min_\lambda \hat{m}_0(\lambda)$ is used as the plug-in estimator of m_0 in the bootstrapping procedure.

Rather than using a bootstrap approach to select an appropriate λ, the HB estimator uses an analogous bootstrap approach to select the number bins for the histogram estimator. This led to better overall performance with respect to RMSE and a lesser degree of negative bias when compared to the LB approach. In particular the estimated RMSE of the HB estimator was lower than the estimated RMSE of the LB estimator for 22 of 30 conditions and exhibited less negative bias in all cases.

Although the bootstrapping approaches performed well when a large proportion of genes were differentially expressed with small to moderate effect sizes, the H20 estimator

proposed in this article was strongest among all estimators when the effect sizes for differentially expressed genes tended to be large. In these situations other approaches tended to suffer from negative bias while the H20 estimator exhibited little bias.

The CD estimator exhibited relatively consistent performance throughout the simulation and was often the top performer among all the estimators with respect to RMSE. The CD estimator seemed particularly good relative to the other methods for the CSY correlation structure although all methods performed somewhat poorly here relative to the other more realistic correlation structures. As expected, all methods tended to perform best when p values were independent, but the methods also performed reasonably well for the AR1 structure which was intended to mimic the structures most often encountered in practice.

6. CONCLUSIONS AND FUTURE DIRECTIONS

The iterative procedure proposed by Mosig et al. (2001) and developed in this article as a noniterative procedure offers a simple and relatively effective method for estimating the number of true null hypotheses when conducting many tests. In our simulation study, the histogram-based estimator with 20 bins tended to be conservative for independent and autoregressive correlation structures in that it tended to overestimate the number of true null hypotheses for the conditions considered in the simulation study of Section 5. The estimator's positive bias tended to diminish as the power for detecting departures from the null increased and as the proportion of true nulls increased. When all null hypotheses were true, the estimator exhibited a slight negative bias. (Note that all reasonable and nontrivial procedures for estimating the number of true null hypotheses must exhibit some negative bias when all nulls are true.) Overall the performance of the 20-bin estimator with respect to estimated bias and mean squared error was very competitive with existing approaches. However, the CD estimator proposed by Langaas et al. (2005) outperformed the 20-bin estimator for 21 of the 30 simulations settings considered in this article and generally exhibited the top performance among all the estimators we considered.

Unfortunately we were unable to consider all possible methods for estimation of the number of true null hypotheses in our simulation study. The method proposed by Allison et al. (2002) is one excluded competitor. Allison et al. (2002) approximated the distribution of p values from multiple tests as an independent and identically distributed sample from a mixture of a uniform and a beta distribution. They used numerical methods to obtain maximum likelihood estimates of the mixing proportions and the parameters of the beta distribution. These estimates can be used to approximate a variety of quantities relevant in a multiple testing setting. In particular, the estimated mixing proportion for the uniform component of the mixture distribution serves as an estimate of π_0. We encountered convergence problems with the available R implementation of this approach when we attempted to include it in our simulation study. In the analysis of individual datasets, simple plots of the estimated mixture distribution overlaid on a histogram of p values can be used to identify problems with convergence. A good fit can often be obtained by trying several starting values for the numerical optimization procedure. However, such a trial-and-error

approach is unwieldy for simulation purposes; thus, we chose not to include this method in our simulation study.

In Section 2.5 we presented a real dataset where the histogram based estimator $\tilde{m}_0(B)$ was relatively insensitive to the number of bins B. In Section 4 we illustrated that the number of bins offers a grid of potential λ values from which to choose a λ-estimator of m_0 as a function of the observed data. In simulations not reported here, we found that the bias of $\tilde{m}_0(B)$ tended to be U-shaped while the variance of $\tilde{m}_0(B)$ tended to have an inverted U shape as B ranged from 2 to 20. In addition, we have studied the performance of the $\tilde{m}_0(B)$ across varying m and B and have found that $\tilde{m}_0(20)$ is a reasonable choice for m as low as 100, yielding RMSE values in the single digits for the situations we studied. Thus, we recommend $\tilde{m}_0(20)$ as a relatively stable estimator of m_0 that typically avoids undesirable negative bias and performed well in the simulation conditions that we considered. Nonetheless, we have offered an alternative bootstrap approach that allows for the number of bins to be selected as a function of the data. This method also performed well in our simulations, exhibiting estimated MSE less than that of the 20-bin estimator for 14 of the 30 simulation conditions. Due to its simplicity and general bias characteristics, we have a slight preference for the 20-bin estimator.

The histogram-based estimator featured in this article appears to be a good candidate for use in the estimation of false discovery rate. The estimator tended to exhibit more positive bias and lower mean squared error than the estimator proposed by Storey and Tibshirani (2003) for most conditions considered in our simulation study. Furthermore, the histogram-based estimator had considerably lower mean squared error than the lowest slope estimator used by Benjamini and Hochberg (2000) while typically maintaining a positive bias when $\pi_0 < 1$ and correlation structures were similar to those expected in practice. This suggests the possibility of more stable estimates of FDR than the Storey and Tibshirani procedure and less conservative estimates of FDR than those provided by the Benjamini and Hochberg (1995, 2000) procedures. Mosig et al. (2001) first proposed the estimator as a means of determining an "adjusted false discovery rate," and Fernando et al. (2004) studied the performance of this estimator in the context of mapping quantitative trait loci.

In future work, we intend to investigate the performance of the histogram-based estimator for computing q values to estimate false discovery rates as described by Storey (2003) and Storey and Tibshirani (2003). Storey, Taylor, and Siegmund (2004) have established theoretical results on FDR control for methods that use $\hat{m}_0(\lambda)$ to estimate m_0. Their results apply directly to the case of a fixed value of λ and can be extended to encompass the case of a randomly selected λ. Given the relationship between the histogram-based estimator and the λ-estimator described in Section 4, it should be possible to obtain results on FDR control for a procedure which uses the histogram-based estimator to estimate m_0.

APPENDIX: PROOF OF THE CONVERGENCE THEOREM

The proof of the convergence theorem rests on the following facts which we state here and prove at the end of the Appendix.

A. $\bar{n}_{1:B} > \bar{n}_{2:B} > \cdots > \bar{n}_{I:B}$.

B. If $i_k < I$, then (i) $i_k \le i_{k+1}$, (ii) $i_{k+1} \le I$, and (iii) $i_k < i_{k^*}$ for some $k^* > k$.

C. If there exists k^* such that $i_{k^*} = I$, then $i_k = I$ for all $k \ge k^*$.

D. Suppose $\{a_k\}_{k\ge 0}$ is an infinite sequence of real numbers. If there exists $\lambda \in [0, 1)$, an integer k^*, and a real number a such that $a_k = \lambda a_{k-1} + (1-\lambda)a$ whenever $k \ge k^*$, then $\lim_{k\to\infty} a_k = a$.

We begin by showing that the sequence $\{i_k\}$ converges to I in a finite number of iterations. Note that if $I = 1$, then $n_1 \le \bar{n}_{1:B} = N_0/B$. Thus $i_1 = 1 = I$ by the definition of i_1, and fact C guarantees that $i_k = I$ for all $i \ge 1$. If $I > 1$, then $i_1 \le I$ because $n_I \le \bar{n}_{I:B} < \bar{n}_{1:B} = N_0/B$ by fact A and the definitions of i_1 and I. Now if $i_1 = I$, fact C guarantees that $i_k = I$ for all $i \ge 1$. If $i_1 < I$, parts (i) through (iii) of fact B imply that i_{k^*} must equal I for some $k^* > 1$. Hence $i_k = I$ for all $k \ge k^*$ by fact C. Thus, in all cases, there is some $k^* \ge 1$ satisfying $i_k = I$ for any $k \ge k^*$. It follows from (2.1) that

$$\frac{N_k}{B} = \left(\frac{I-1}{B}\right)\frac{N_{k-1}}{B} + \left(1 - \frac{I-1}{B}\right)\bar{n}_{I:B} \quad \text{for all} \quad k \ge k^*.$$

Now fact D implies that $\lim_{k\to\infty} \frac{N_k}{B} = \bar{n}_{I:B}$, and the desired result follows. □

Proof of Fact A: For all $i < I$ we have

$$n_i > \bar{n}_{i:B} = \left(\frac{1}{B-i+1}\right) n_i + \left(1 - \frac{1}{B-i+1}\right)\bar{n}_{i+1:B} \tag{A.1}$$

Straightforward manipulation of (A.1) yields $n_i > \bar{n}_{i+1:B}$. Now substituting $\bar{n}_{i+1:B}$ for n_i in the right side of equation (A.1) yields $\bar{n}_{i:B} > \bar{n}_{i+1:B}$. Thus,

$$\bar{n}_{1:B} > \bar{n}_{2:B} > \cdots > \bar{n}_{I:B}.$$

□

Proof of Fact B(i): By the definition of I, $n_{i_k} > \bar{n}_{i_k:B}$ whenever $i_k < I$. By the definition of i_k, $n_{i_k} \le N_{k-1}/B$. Thus $\bar{n}_{i_k:B} < N_{k-1}/B$. Substituting N_{k-1}/B for $\bar{n}_{i_k:B}$ in (2.1) yields

$$\frac{N_k}{B} < \left(\frac{i_k-1}{B}\right)\frac{N_{k-1}}{B} + \left(1 - \frac{i_k-1}{B}\right)\frac{N_{k-1}}{B} = \frac{N_{k-1}}{B}.$$

Therefore $n_i \le N_k/B$ implies that $n_i < N_{k-1}/B$. It follows that $\{i : n_i \le N_k/B\} \subseteq \{i : n_i \le N_{k-1}/B\}$. Thus $i_k = \min\{i : n_i \le N_{k-1}/B\} \le \min\{i : n_i \le N_k/B\} = i_{k+1}$. □

Proof of Fact B(ii): By definition of i_k, $n_{i_k} \le N_{k-1}/B$. Thus, using fact A and the definition of I, we have

$$n_I \le \bar{n}_{I:B} \le \bar{n}_{i_k:B} < n_{i_k} \le N_{k-1}/B$$

whenever $i_k < I$. Thus

$$n_I < \left(\frac{i_k-1}{B}\right)\frac{N_{k-1}}{B} + \left(1 - \frac{i_k-1}{B}\right)\bar{n}_{i_k:B} = \frac{N_k}{B}$$

by (2.1) which implies that $i_{k+1} \leq I$ by the definition of i_{k+1}. □

Proof of Fact B(iii): Suppose $i_k < I$ and $i_\ell \leq i_k$ for all $\ell > k$. This would imply that $i_\ell = i_k$ for all $\ell \geq k$ by fact B(i). Thus

$$\frac{N_\ell}{B} = \left(\frac{i_k - 1}{B}\right)\frac{N_{\ell-1}}{B} + \left(1 - \frac{i_k - 1}{B}\right)\bar{n}_{i_k:B} \quad \text{for all} \quad \ell \geq k.$$

Fact D implies that $\lim_{\ell\to\infty} N_\ell/B = \bar{n}_{i_k:B}$. However, $i_k < I$ implies that $n_{i_k} > \bar{n}_{i_k:B}$. Thus, there exists $k^* > k$ such that $n_{i_k} > N_{\ell-1}/B$ whenever $\ell \geq k^*$. This implies that $i_\ell \neq i_k$ when $\ell \geq k^*$. We have reached a contradiction. Therefore the result follows. □

Proof of Fact C: If $i_k = I$, then (2.1) implies that

$$\frac{N_k}{B} = \left(\frac{I - 1}{B}\right)\frac{N_{k-1}}{B} + \left(1 - \frac{I - 1}{B}\right)\bar{n}_{I:B}. \tag{A.2}$$

By the definition of I, $n_I \leq \bar{n}_{I:B}$. By the definition of i_k, $n_I \leq N_{k-1}/B$ when $i_k = I$. Thus, substituting n_I for N_{k-1}/B and n_I for $\bar{n}_{I:B}$ in (A.2) yields $N_k/B \geq n_I$. Hence $i_{k+1} \leq I$. Now by fact A and the definition of I, $n_i > \bar{n}_{I:B}$ for any $i < I$. Furthermore $i_k = I$ implies $n_i > N_{k-1}/B$ for any $i < I$. Thus substituting n_i for $\bar{n}_{I:B}$ and n_i for N_{k-1}/B in (A.2) implies $n_i > N_k/B$ for any $i < I$. Thus, $i_{k+1} \geq I$. We have shown that $i_{k+1} \leq I$ and $i_{k+1} \geq I$. Thus, $i_{k+1} = I$. □

Proof of Fact D: Let $b_n = a_{k^*+n-1}$ for all $n \geq 0$. Then $b_1 = \lambda b_0 + (1 - \lambda)a$, and an induction argument shows that $b_n = \lambda^n b_0 + a(1 - \lambda)\sum_{i=0}^{n-1} \lambda^i$ for all $n \geq 1$. Now $\lambda \in [0, 1)$ implies that

$$\lim_{n\to\infty} \lambda^n = 0 \quad \text{and} \quad \sum_{i=0}^{\infty} \lambda^i = (1 - \lambda)^{-1}.$$

Thus, $\lim_{n\to\infty} b_n = a$, and $\lim_{k\to\infty} a_k = a$ since $\{a_k\}_{k\geq k^*} = \{b_n\}_{n\geq 1}$. □

ACKNOWLEDGMENTS

The authors thank Gary Gadbury, Mette Langaas, and John Storey for providing R code for three of the estimators considered in this article. D. Nettleton acknowledges support of the Cooperative State Research, Education, and Extension Service, U.S. Department of Agriculture, under Agreement No. 2002-35300-12619.

[Received August 2005. Revised March 2006.]

REFERENCES

Allison, D. B., Gadbury, G. L., Heo, M., Fernández, J. R., Lee, C.-K., Prolla, T. A., and Weindruch, R. (2002). "A Mixture Model Approach for the Analysis of Microarray Gene Expression Data," *Computational Statistics and Data Analysis*, 39, 1–20.

Benjamini, Y., and Hochberg, Y. (1995), "Controlling False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," *Journal of the Royal Statistical Society*, Series B, 57, 289–300.

——— (2000), "On the Adaptive Control of the False Discovery Rate in Multiple Testing with Independent Statistics," *Journal of Educational and Behavioral Statistics*, 25, 60–83.

Brem, R. B., Yvert, G., Clinton, R., and Kruglyak, L. (2002), "Genetic Dissection of Transcriptional Regulation in Budding Yeast," *Science*, 296, 752–755.

Bystrykh, L., Weersing, E., Dontje, B., Sutton, S., Pletcher, M. T., Wiltshire, T., Su, A. I., Vellenga, E., Wang, J., Manly, K. F., Lu, L., Chesler, E. J., Alberts, R., Jansen, R. C., Williams, R. W., Cooke, M. P. and de Haan, G. (2005). "Uncovering Regulatory Pathways that Affect Hematopoietic Stem Cell Function Using 'Genetical Genomics'," *Nature Genetics*, 37, 225–232.

Caldo, R. A., Nettleton, D., and Wise, R. P. (2004), "Interaction-Dependent Gene Expression in Mla-Specified Response to Barley Powdery Mildew," *The Plant Cell*, 16, 2514–2528.

Chesler, E. J., Lu, L., Shou, S., Qu, Y., Gu, J., Wang, J., Hsu, H. C., Mountz, J. D., Baldwin, N. E., Langston, M. A., Threadgill, D. W., Manly, K. F. and Williams, R. W. (2005), "Complex Trait Analysis of Gene Expression Uncovers Polygenic and Pleiotropic Networks that Modulate Nervous System Function," *Nature Genetics*, 37, 233–242.

Close, T. J., Wanamaker, S., Caldo, R., Turner, S. M., Ashlock, D. A., Dickerson, J. A., Wing, R. A., Muehlbauer, G. J., Kleinhofs, A. and Wise, R. P. (2004), "A New Resource for Cereal Genomics: 22K Barley GeneChip Comes of Age," *Plant Physiology*, 134, 960–968.

DeCook, R., Lall, S., Nettleton, D., and Howell, S. H. (2006), "Genetic Regulation of Gene Expression During Shoot Development in Arabidopsis," *Genetics*, 172, 1155–1164.

Fernando, R. L., Nettleton, D., Southey, B. R., Dekkers, J. C. M., Rothschild, M. F., and Soller, M. (2004), "Controlling the Proportion of False Positives (PFP) in Multiple Dependent Tests," *Genetics*, 166, 611–619.

Genovese, C. R., and Wasserman, L. (2004), "A Stochastic Process Approach to False Discovery Control," *The Annals of Statistics*, 32, 1035–1061.

Hochberg, Y., and Benjamini, Y. (1990), "More Powerful Procedures for Multiple Significance Testing," *Statistics and Medicine*, 9, 811–818.

Hsueh, H., Chen, J. J., and Kodell, R. L. (2003), "Comparison of Methods for Estimating the Number of True Null Hypotheses in Multiplicity Testing," *Journal of Biopharmaceutical Statistics*, 13, 675–689.

Hubner, N., Wallace, C. A., Zimdahl, H., Petretto, E., Schulz, H., Maciver, F., Mueller, M., Hummel, O., Monti, J., Zidek, V., Musilova, A., Kren, V., Causton, H., Game, L., Born, G., Schmidt, S., Müller, A., Cook, S., Kurtz, T. W., Whittaker, J., Pravenec, M., and Aitman, T. J. (2005), "Integrated Transcriptional Profiling and Linkage Analysis for Identification of Genes Underlying Disease," *Nature Genetics*, 37, 243–253.

Jansen, R. C., and Nap, J. P. (2001), "Genetical Genomics: The Added Value from Segregation," *Trends in Genetics*, 17, 388–391.

Langaas, M., Ferkingstad, E., and Lindqvist, B. H. (2005), "Estimating the Proportion of True Null Hypotheses, with Application to DNA Microarray Data," *Journal of the Royal Statistics Society*, Series B, 67, 555–572.

Lipshutz, R. J., Fodor, S. P., Gingeras, T. R. and Lockhart, D. J. (1999), "High Density Synthetic Oligonucleotide Arrays," *Nature Genetics*, 21 Supplement, 20–24.

Mosig, M. O., Lipkin, E., Galina, K., Tchourzyna, E., Soller, M., and Friedmann, A. (2001), "A Whole Genome Scan for Quantitative Trait Loci Affecting Milk Protein Percentage in Israeli-Holstein Cattle, by Means of Selective Milk DNA Pooling in a Daughter Design, Using an Adjusted False Discovery Rate Criterion," *Genetics*, 157, 1683–1698.

Nguyen, D. V. (2004), "On Estimating the Proportion of True Null Hypotheses for False Discovery Rate Controlling Procedures in Exploratory DNA Microarray Studies," *Computational Statistics and Data Analysis*, 47, 611–637.

Pomp, D., Allan, M. F., and Wesolowski, S. R. (2004), "Quantitative Genomics: Exploring the Genetic Architecture of Complex Trait Predisposition," *Journal of Animal Science*, 82, E300–312.

Schadt, E.E., Monks, S.A., Drake, T.A., Lusis, A.J., Che, N., Colinayo, V. Ruff, T.G., Milligan, S.B., Lamb, J.R., Cavet, G., Linsley, P.S., Mao, M., Stoughton, R.B., and Friend, S.H. (2003a), "Genetics Of Gene Expression

Surveyed In Maize, Mouse And Man," *Nature*, 422, 297–302.

Schadt, E. E., Monks, S. A., and Friend, S. H. (2003b), "A New Paradigm for Drug Discovery: Integrating Clinical, Genetic, Genomic and Molecular Phenotype Data to Identify Drug Targets," *Biochemical Society Transactions*, 31, 437–443.

Schweder, T., and Spjøtvoll, E. (1982), "Plots of P-values to Evaluate Many Tests Simultaneously," *Biometrika*, 69, 493–502.

Simes, R. J. (1986), "An Improved Bonferroni Procedure for Multiple Tests of Significance," *Biometrika*, 73, 751–754.

Storey, J. D. (2002a), "A Direct Approach to False Discovery Rates," *Journal of the Royal Statistical Society*, Series B, 64, 479–498.

——— (2002b), "False Discovery Rates: Theory and Applicatons to DNA Microarrays," unpublished Ph.D. thesis, Department of Statistics, Stanford University.

——— (2003), "The Positive False Discovery Rate: A Bayesian Interpretation and the q-Value," *The Annals of Statistics*, 31, 2013–2035.

Storey, J. D., Taylor, J. E., and Siegmund, D. (2004), "Strong Control, Conservative Point Estimation, and Simultaneous Conservative Consistency of False Discovery Rates: A Unified Approach," *Journal of the Royal Statistical Society*, Series B, 66, 187–205.

Storey, J. D., and Tibshirani, R. (2003), "Statistical Significance for Genomewide Studies," in *Proceedings of the National Academy of Sciences*, 100, pp. 9440–9445.

Yvert, G., Brem, R.B., Whittle, J., Akey, J.M., Foss, E., Smith, E.N., Mackelprang, R., and Kruglyak, L. (2003), "*Trans*-acting Regulatory Variation in *Saccharomyces cerevisiae* and the Role of Transcription Factors," *Nature Genetics*, 35, 57–64.