2011

Generalized Benjamini-Hochberg procedures using spacings

Debashis Ghosh, *Penn State University*

Generalized Benjamini-Hochberg procedures using spacings

Debashis Ghosh

Departments of Statistics and Public Health Sciences, Pennsylvania State University, University Park, PA, USA

Summary. For the problem of multiple testing, the Benjamini-Hochberg (B-H) procedure has become a very popular method in applications. We show how the B-H procedure can be interpreted as a test based on the spacings corresponding to the p-value distributions. Using this equivalence, we develop a class of generalized B-H procedures that maintain control of the false discovery rate in finite-samples. We also consider the effect of correlation on the procedure; simulation studies are used to illustrate the methodology.

1. Introduction

Consideration of high-dimensional data has become the norm in applied statistical practice these days. For example, in genomics, it is quite common to look for genes that are up- or down-regulated in cancerous tissue relative to healthy tissue using high-throughput data (e.g., microarrays or next-generation sequencing technologies). Similarly, in neuroimaging, there is consideration of thousands of voxels as a global map of activity in the human brain using functional magnetic resonance imaging technology. Assessing differential expression in these settings leads to a massive multiple comparisons problem that results from performing thousands of tests for each gene or voxel. This has led to an explosion of literature on statistical methods for multiple testing. Many authors have recently advocated for control of the false discovery rate (FDR) (Benjamini and Hochberg, 1995) relative to the traditional familywise type I error (FWER). Many authors have studied and developed methods for controlling the false discovery rate (e.g., Efron et al., 2001; Sarkar, 2002; Efron, 2004; Storey et al., 2004; Genovese and Wasserman, 2002, 2004; Cohen and Sackrowitz, 2005; Lehmann and Romano, 2005; Sarkar, 2006; Ferreira and Zwinderman, 2006; Jin and Cai, 2007; Chi, 2007; Sarkar and Guo, 2009; Finners et al., 2009). The above list is far from being completely exhaustive.

The Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995) has received much recent attention in that it controls FDR and can lead to greater power relative to a multiple testing adjustment using the Bonferroni correction, for example. Much recent work has focused on adaptive versions of the B-H procedure (e.g., Benjamini et al., 2006); in this literature, one attempts to estimate the proportion of true null hypotheses and adjust the threshold in the Benjamini-Hochberg procedure accordingly. In addition, many authors have noted the equivalence between the B-H procedure with thresholding based on the empirical distribution of the p-values, which has allowed for development of theoretical

results using empirical process theory and related techniques (Storey et al., 2004; Genovese and Wasserman, 2004; Ferreira and Zwinderman, 2006).

Our starting point is quite different from the above. In particular, we make explicit use of the fact that the original B-H procedure involved *sorting* of the p-values in increasing order. This sorting operation can be equivalently characterized in terms of spacings (Pyke, 1965). Doing so leads to a new characterization of the B-H procedure as well as some new extensions that are FDR-controlling procedures. Both exact and asymptotic results will be presented. A major implication of our procedures is that for multiple testing, it may not be the case that evidence for rejecting null hypotheses exist solely in the left-tail of the distribution of the p-values. The structure of this paper is as follows. In Section 2, we give some background on multiple testing and discuss the procedure of Benjamini and Hochberg (1995) as well as related work. In Section 3, we recast the B-H procedure using spacing results and propose a so-called generalized B-H procedures. We demonstrate FDR control in finite-samples in Section 4. Section 5 explores the effects of correlation on the proposed procedures. Some simulation studies illustrating the proposed methodology are given in Section 6. We conclude with some discussion in Section 7.

2. Multiple Testing Background

Suppose we have test statistics $T_1, \ldots, T_n$ for testing hypotheses H_{0i}, $i = 1, \ldots$,n. Suppose we are interested in testing a set of n hypotheses. Of these n hypotheses, suppose that for n_0 of them, the null is true. The FDR can be conceptualized using the following 2×2 contingency table:

[Note: Table 1 about here.]

The definition of false discovery rate (FDR) as put forward by Benjamini and Hochberg (1995) is

$$FDR \equiv E\left[\frac{V}{Q} \mid Q > 0\right] P(Q > 0).$$

The conditioning on the event $[Q > 0]$ is needed because the fraction V/Q is not well-defined when $Q = 0$. Benjamini and Hochberg (1995) propose a simple algorithm for selecting the hypotheses that are significant that controls the false discovery rate (FDR). Note that it involves converting the statistics $T_1, \ldots, T_n$ into p-values $p_1, \ldots, p_n$. Note that there are many ways in which this could be done, such as model-based p-values or permutation methods. In this paper, all we will assume is that there exists some method of converting test statistics into p-values. Let α denote the rate at which it is desired to control the false discovery rate. The algorithm of Benjamini and Hochberg (1995) is then summarized in Box 1.

Box 1. Benjamini and Hochberg (1995) procedure

(a) Let $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(n)}$ denote the ordered, observed p-values.

(b) Find $\hat{k} = \max\{1 \leq i \leq n : p_{(i)} \leq \alpha i/n\}$.

(c) If $\hat{k}$ exists, then reject null hypotheses $p_{(1)} \leq \cdots \leq p_{(\hat{k})}$. Otherwise, reject nothing.

Benjamini and Hochberg (1995) show that the procedure in Box 1 controls the FDR at level α when the p-values are independent and uniformly distributed. Benjamini and Yekutieli (2001) show that the procedure in Box 1 controls the FDR at level α under the condition that the joint distribution of the test statistics corresponding to the true null hypotheses are positively regression dependent. A more recent simplified proof of the FDR control was developed by Finners et al. (2009). We also mention that FDR is the expected value of the so-called false discovery proportion (FDP).

In terms of results involve error control of the FDR procedure, there are two types of results typically found: exact FDR control in finite samples and asymptotic FDR control. For the first class of results, Benjamini and Yekutieli (2001) showed that the Benjamini-Hochberg procedure has exact error control under positive regression dependence. A simplified proof was given by Storey et al. (2004) in the situation when the p-values for the true null hypotheses are independent. More recently, Sarkar (2002, 2006) has developed useful inequalities in order to prove the exact FDR control of generalizations of the Benjamini-Hochberg procedure under positive regression dependence. Sarkar et al. (2008) proved the exact FDR control of a procedure based on the so-called Bayesian false discovery rate.

With respect to asymptotic control of FDR, authors such as Storey et al. (2004), Genovese and Wasserman (2004) and Ferreira and Zwinderman (2006) have all observed the following fact: the B-H procedure is equivalent to the thresholding rule $c_\alpha(\widehat{FDR})$, where

$$c_\alpha(F) = \sup\{0 \leq t \leq 1 : F(t) \leq \alpha\},$$

where $\widehat{FDR}$ is an estimator of the false discovery rate. Using this equivalence, the previously mentioned authors are able to derive asymptotic results. In their work, the necessary crucial assumption is that the distribution functions for the p-values corresponding to the true nulls and true alternatives converge to population limits. This is a Glivenko-Cantelli type of result, so the results that have been derived can allow for some dependence between the p-values.

3. Spacings and the B-H procedure

Spacings have a long history in statistics, dating back to the first half of the 20th century (e.g., Pearson, 1902; Greenwood, 1946; Moran, 1947). A very comprehensive review of the topic can be found in Pyke (1965). Let $U_1, \ldots, U_n$ denote a random sample from the Uniform(0,1) distribution. Then the spacings are defined as $V_i = U_{(i)} - U_{(i-1)}$, for

$i = 1, \ldots, n+1$, where $U_{(i)}$ denotes the ith order statistic of U, $U_{(0)} = 0$ and $U_{(n+1)} = 1$. Note that while the joint density of $U_1, \ldots, U_n$ is

$$f(u_1, \ldots, u_n) = \begin{cases} 1 \text{ if } 0 \leq u_i \leq 1 \text{ for all } i \\ 0 \text{ otherwise,} \end{cases}$$

the joint density of the spacings is given by

$$f(v_1, \ldots, v_n) = \begin{cases} n! \text{ if } v_i \geq 0 \text{ for all } i \text{ and } \sum_{i=1}^{n+1} v_i = 1 \\ 0 \text{ otherwise.} \end{cases} \tag{1}$$

We observe that the joint density of the spacings is defined on the simplex $\{\sum_{i=1}^{n+1} v_i = 1, v_i \geq 0, i = 1, \ldots, n+1\}$. However, we also note that the V's are not statistical independent because of the constraint that they must sum to one. Similarly, the order statistics are not statistically independent even if the original random variables are independent and identically distributed.

Based on (1), we observe that the joint density of the spacings is invariant to permutation of the indices. This implies that the marginal distribution of V_i will be the same as that of V_1. It can be shown that $E(V_i) = (n+1)^{-1}$ for $i = 1, \ldots, n+1$ and that $E(V_j V_k) = (n+1)^{-1}(n+2)^{-1}$ for $1 \leq j, k \leq n+1$. From these expressions, it is easy to find the second moment of V as well as its variance. Determining higher-order moments of V can also be done, although it becomes much more algebraically tedious. In fact, because the U's are distributed from a Uniform(0,1) distribution, we can say that marginally, the V_i's will have a Beta(1,n) distribution, i.e., the pdf of V is given by

$$f(v) = n(1-v)^{n-1}.$$

However, we again stress that there is dependence among $V_1, \ldots, V_n$.

We now return to the Benjamini-Hochberg procedure. We will begin by assuming that the p-values $p_1, \ldots, p_n$ are statistically independent. Let us define $p_{(0)} = 0$, $p_{(n+1)} = 1$ and

$$\tilde{p}_i = p_{(i)} - p_{(i-1)}, \qquad i = 1, \ldots, n+1.$$

It is straightforward to show that we can express $p_{(i)} = \sum_{j=1}^{i} \tilde{p}_i$. Based on this and the fact that $E(\tilde{p}_1) = (n+1)^{-1}$, we can express the B-H procedure from Box 1 in the following manner: reject $p_{(1)}, \ldots, p_{(\hat{k})}$, where

$$\hat{k} = \max\{i : i^{-1} \sum_{j=1}^{i} \tilde{p}_i \leq \alpha(n+1)n^{-1}E(\tilde{p}_1)\}, \tag{2}$$

with $\hat{k} = 0$ if the set in (2) is empty. Equation (2) is illuminating. It says that the B-H procedure can be phrased in terms of a comparison between the cumulative average of the spacings with the corresponding expected value. As demonstrated in Benjamini and Hochberg (1995), such a procedure has expected value $n_0\alpha/n$, where α is the desired FDR

to control. We can also view (2) as a test for clustering of p-values on the unit interval, which is one application of spacings in statistics (e.g., Moran, 1947; Cressie, 1979). The B-H procedure assesses clustering of p-values using first-order differences in the sorted p-values. In particular, gaps between the sorted p-values that are smaller than expected, i.e. $E(\tilde{p}_1)$, constitute evidence against the null hypothesis. As will be seen in the next section, we can make other choices for assessing clustering of p-values in the multiple testing problem that will also lead to FDR-controlling procedures. Finally, we also observe the presence of the factor $(n+1)/n$ in (2). We note that as n tends to infinity, this factor will approach one.

4. Proposed methodology: exact results

Based on analogy with the B-H procedure, we can extend it in an obvious way using first-order differences (i.e. the spacings). This leads to what we term the generalized Benjamini-Hochberg (gBH) procedure.

gBH procedure: Reject $p_{(1)}, \ldots, p_{(\hat{k})}$, where

$$\hat{k} = \max[i : i^{-1} \sum_{j=1}^{i} g(\tilde{p}_j) \leq \alpha E\{g(\tilde{p}_1)\}], \tag{3}$$

g is some suitably chosen monotonic function, and $\hat{k} = 0$ if the set in (3) is empty.

If g in (3) is taken to be the identity function, then this will almost yield the original B-H procedure. The difference will be the factor $(n+1)/n$ on the right-hand side of the equality sign. However, if we replace $i\alpha/n$ by $i\alpha/(n+1)$ in part (b) of Box 1, then this is in fact slightly more stringent than the B-H procedure so that we will still maintain FDR control.

For the procedure in (3) to work in practice, it is desirable to choose a g function so that the expected value in the expression is analytically tractable. A natural class of functions to use is the following family of power functions: for $\lambda \geq 0$

$$g_\lambda(z) = \begin{cases} z^\lambda & \lambda > 0, \\ \log(z) & \lambda = 0. \end{cases} \tag{4}$$

By exploiting the fact that $\tilde{p}_1$ has a Beta distribution, it is easy to show that

$$E\{g_\lambda(\tilde{p}_1)\} = \begin{cases} B(1+\lambda, n+1)/B(1, n+1) & \lambda > 0, \\ \psi(1) - \psi(n+2) & \lambda = 0 \end{cases},$$

where $B(u, v) = \Gamma(u)\Gamma(v)[\Gamma(u+v)]^{-1}$, and

$$\psi(x) \equiv \frac{d}{dx} \log \Gamma(x) = \frac{\Gamma'(x)}{\Gamma(x)}$$

is the digamma function.

One important question to answer is whether or not the proposed procedure maintains the proper error control. To this end, we have the following theorem:

Theorem 1: *Assuming that the p-values are independent, the generalized B-H procedure given by (3) maintains FDR control at level* α.

Proof: Define the filtration $\mathcal{F}_k = \sigma(\{\tilde{p}_i, n-k+1 \leq i \leq n\})$, the $\sigma-$field generated by the first k spacings of the p-values in reverse-time. Then by basic properties of order statistics (Pyke, 1965, p. 399), the random variable

$$V_k \equiv k^{-1} \sum_{i=n-k+1}^{n} [g(\tilde{p}_i) - E\{g(\tilde{p}_1)\}]$$

is a zero-mean martingale with respect to $\mathcal{F}_k$. Then $\hat{k}$ in (3) represents a stopping time with respect to the filtration so by the optional sampling theorem,

$$E(V_k|\mathcal{F}_{\hat{k}}) = V_{\hat{k}} \leq \alpha.$$

Remark 1. The proof of Theorem 1 implicitly utilizes the fact that the spacings between the order statistics of the p-values are statistically independent. In fact, the result would still hold if V_k were a supermartingale. We will explore robustness of the methodology to dependence in Section 5.

Remark 2. The B-H procedure as originally conceived uses (4), where $\lambda = 1$. The only type of information that is being used is the first-order differences between the sorted p-values for assessing clustering of p-values. However, other values could be used as well. For instance, the value of $\lambda = 2$ has been argued as an optimal value for assessing clustering on theoretical grounds by Cressie (1979).

Remark 3. Given that the spacings approach has led to a reinterpretation of the multiple testing problem as one of assessing clustering against a uniform distribution, this leads to a fundamental philosophical discrepancy for multiple comparisons. To illustrate the issue, we perform a simple simulation. We generate $n = 10000$ p-values as a random sample from the following mixture model:

$$p_1, \ldots, p_n \sim 0.9U(0,1) + 0.05\text{Beta}(3,30) + 0.05\text{Beta}(30,3),$$

where $\text{Beta}(\alpha, \beta)$ denotes the pdf

$$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}.$$

The picture is given in Figure 1. Based on the picture, we have evidence against uniformly distributed spacings based on both small p-values as well as big p-values. Traditionally, most researchers have ruled out evidence being contained in the right-tail of the distribution in Figure 1. This has been done in two ways. One has been to assume that the distribution of p-values come from a two-component mixture model in which the second component,

which represents p-values from the alternative, is stochastically smaller than a Uniform(0,1) random variable. A second way has been to assume that the distribution function of p-values across all hypotheses is concave. However, if these conditions do not hold, then it could be the case that **large** p-values are the ones that are providing evidence against the null hypothesis, as is the case in Figure 1. By treating the hypothesis testing problem as one of spatial clustering against the null hypothesis of spacings corresponding to statistics that are uniformly distributed on $[0, 1]$, this allows for information from both tails of the distribution in Figure 1 to contribute to evidence against the null hypothesis. More generally, while univariate p-values are important for assessing the evidence of individual hypotheses, for the multiple testing problem, what is more relevant is the aggregate behavior of the ensemble of p-values relative to a reference distribution. The Benjamini-Hochberg procedure uses the Uniform(0,1), but other choices may be possible (e.g., Efron, 2004). Other criticisms of p-value based methods for multiple testing have been given by Cohen and Sackrowitz (2005), Jin and Cai (2007) and Chi (2007).

Remark 4. It is of interest to compare the proposed methodology with the approach considered by Jin and Cai (2007). In particular, they treat the multiple problem by consideration of an optimization problem based on a normal mixture model. Their methodology leads to the optimality of the local false discovery rate for multiple testing. Their approach is not directly applicable to the setting here, as they work with test statistics rather than p-values. To simplify the exposition, we assume that the p-values come from a mixture of two distributions,

$$p_1, \ldots, p_n \overset{iid}{\sim} \pi_0 U(0,1) + (1-\pi_0) F_U, \tag{5}$$

where F_U is assumed be the cumulative distribution function for a random variable that is stochastically smaller than the Uniform(0,1) random variable. The Jin-Cai algorithm involves computing the local false discovery rate (lFDR)

$$lfdr_i = P(\text{no DE}|p_i) = \frac{\pi_0}{f(p_i)}, \qquad i = 1, \ldots, n,$$

where $f(p)$ is the density of p_i. Based on the mixture model (5), the local false discovery rate defined above is simply the posterior probability that $p-$value i is from the uniform (0,1) component. The Jin-Cai algorithm then rejects hypotheses $H_{(1)}, \ldots, H_{(\hat{J})}$, where

$$\hat{J} = \max\{i : i^{-1} \sum_{k=1}^{i} \widehat{lfdr_{(k)}} \leq \alpha\},$$

where $\widehat{lfdr_{(i)}}$ are the order statistics based on sample-based estimators of $lfdr_i$, $i = 1, \ldots, n$. As before, if the set is empty, then we define $\hat{J} = 0$. Simply comparing the structure of $\hat{J}$ with (2) implies that if the spacings and local false discovery rates are equal, then the two procedures will reject the same hypotheses. This will not be true in general.

5. Effects of dependence

So far, we assumed that the p-values are a random sample and hence are statistically independent. In most practical applications, this is not a reasonable assumption. In this section, we consider the effects of dependence. Let us reconsider the gBH procedure again. The method compares an empirical average of spacings to its expected value, scaled by the FDR. If there is correlation among the p-values, intuitively we expect that positive correlations between them will make spacings shorter than those from a random sample, while negative correlations will have an opposite effect.

Some progress can be made in the case of positive correlation. To do so requires results from the theory of stochastic majorization (Proschan and Sethuraman, 1977; Nevius et al., 1977). For two vectors $\mathbf{u}$ and $\mathbf{v}$, we define the majorization partial ordering as $\mathbf{u} \prec \mathbf{v}$ if

$$\begin{aligned} \sum_{k=l}^{J} u_{\gamma(k)} &\leq \sum_{k=l}^{J} v_{\beta(k)}, \quad l = 2, \ldots, J \\ \sum_{k=1}^{J} u_{\gamma(k)} &= \sum_{k=1}^{J} v_{\beta(k)}, \end{aligned}$$

where γ and β are permutations of the indices $1, \ldots, J$ that reorder $\mathbf{u}$ and $\mathbf{v}$, i.e.

$$u_{\gamma(1)} \leq u_{\gamma(2)} \leq \cdots \leq u_{\gamma(J)}, \quad v_{\beta(1)} \leq v_{\beta(2)} \leq \cdots \leq v_{\beta(J)}.$$

A function $f : R^J \rightarrow R^n$ is said to be Schur-convex if $\mathbf{u} \prec \mathbf{v}$ implies that $f(\mathbf{u}) \leq f(\mathbf{v})$ componentwise in R^n. A random vector $\mathbf{U}$ stochastically majorizes a random vector $\mathbf{V}$ if $f(\mathbf{U}) \geq_{s.t.} f(\mathbf{V})$ for every Schur-convex function $f : R^J \rightarrow R$, where $X \geq_{s.t.} Y$ denotes that X is stochastically larger than Y or equivalently, that $P(X > x) \geq P(Y > x)$ for all x. Let us define a testing procedure to be monotonic if the following condition holds: if $p_1, \ldots, p_n$ are p-values and $p'_1, \ldots, p'_n$ are another set of p-values with $p'_i \leq p_i$, then $\text{FDP}(p_1, \ldots, p_n) \geq \text{FDP}(p'_1, \ldots, p'_n)$. We have the following result:

Lemma 1: Assuming the monotonicity of FDP, if the spacings corresponding to the joint distribution of $\mathbf{p}$ is stochastically majorized by the joint distribution of spacings for n independent Uniform(0,1) random variables, then the gBH procedure will provide exact control of the FDR.

Proof: Let $\mathbf{q}$ denote the $n-$dimensional vector whose components are a random sample from the Uniform(0,1) distribution. By assumption, the spacings of $\mathbf{p}$ stochastically majorizes the spacings $\mathbf{q}$. Since the sum function is a Schur-convex function, this implies that the cumulative average of spacings for $\mathbf{p}$ will be smaller than the corresponding quantity based on $\mathbf{q}$. Since the gBH procedure controls the FDR for $\mathbf{q}$, because of the monotonicity of FDP, it will also control the FDR for $\mathbf{p}$.

It is now important to develop a characterization of the classes of distributions that will stochastically majorize a random sample of n Uniform(0,1) p-values. We formulate the

following model:

$$Pr(T_1 \leq t_1, T_2 \leq t_2, \ldots, T_n \leq t_n) = C\{Pr(T_1 \leq t_1), Pr(T_2 \leq t_2), \ldots, Pr(T_n \leq t_n)\}, \quad (6)$$

where C is a function that maps from $[0,1]^n$, the n-fold product space of $[0,1]$, to $[0,1]$. Then it can be shown that given the joint distribution of $T_1, \ldots, T_n$, there always exists a function C such that (6) holds with the marginal distribution of T_i being F_i, $i = 1, \ldots, n$. Equation (6) is known as a copula model.

Example 1: (Archimedean Copulas). As mentioned earlier, one particular class of copulas are Archimedean copulas. The copula function defines a proper joint distribution if ϕ refers to the Laplace transform of a nonnegative random variable W. These probability models also have a two-stage formulation. We can write Archimedean copulas in the following manner:

$$C_\theta(u_1, \ldots, u_n) = \int F^\theta(u) dG(\theta), \quad (7)$$

where F are univariate cumulative distribution functions, and G is a mixing distribution for θ. It can be shown using arguments in Ahmed et al. (1981) that this copula function satisfies the positive regression dependency criterion. Note that this class of examples will include random effects models.

6. Simulation Studies

Here, we report on the results of some simulation studies to assess the finite-sample properties of the proposed methodologies. We generated $n = 300$ statistics using a multivariate normal distribution with mean μ_i for the ith statistic and variance one. We set n_1 of the test statistics to have $\mu_i = 2$ and the remaining ones to have mean $\mu_i = 0$. Note that our definition of n_1 corresponds to the true alternatives. Data were simulated assuming a correlation of zero (independent case) and 0.3 (dependent case). Note that for this condition, the positive dependence conditions in Lemma 1 are satisfied. We studied several choices of g in the generalized BH procedure based on the power family (4) with differing values of λ. While not reported here, we found poor performance using $\lambda = 0$. In addition to the false discovery rate, we also calculated one minus the non-discovery rate (NDR). This is defined as the number of false null hypotheses (i.e., statistics generated from the normal distribution with nonzero mean) that are rejected by the testing procedure. Loosely speaking, $1 - NDR$ can be loosely thought of as a measure of the power of the procedure.

The results are given in Tables 2 and 3. All procedures are controlled at an FDR of 0.05, with the exception of the generalized BH procedures with larger values of λ. We find that in terms of FDR, the operating characteristics of the BH procedure and the proposed methodology with $\lambda = 2$ are quite similar in the independent case. The FDR typically tends to be conservative, leaving open the possibility that adaptive methods (e.g., Benjamini et al., 2006) should be explored as well. However, in the case of dependence, the proposed methodology achieves an FDR closer to the nominal one when $\lambda = 2$. In addition, the

proposed procedures enjoy much higher power relative to the B-H procedure; the difference is more pronounced in the dependent case. While these results are quite suggestive, further investigation is needed.

7. Discussion

In this article, we have developed a new family of multiple testing procedures based on generalization of the B-H procedure by incorporating the fact that the B-H procedure can be thought of as a test of clustering based on sample averages of spacings compared to its expected value. This leads to a natural extension of B-H procedures that we have shown to maintain FDR control and k-FDR control in finite samples as well as asymptotically. Simulation studies show that the generalized B-H procedure has suitable operating characteristics.

While the B-H procedure does compare empirical averages of spacings to its expected value uniformity, there is a restriction made that only small p-values can be rejected. This is problematic for a situation such as Figure 1, in which evidence against the null hypothesis is suggested by both small and large p-values. This highlights a fundamental disconnect in the problem of multiple comparisons. One approach would be to use a different reference distribution other than uniform. Such an approach has been taken by Efron (2004); he uses the concept of a theoretical null distribution. However, that work only dealt with the test statistic scale; we are currently exploring extensions to p-values. An alternative approach is to not work with p-values; this is what is done by Sun and Cai (2007). However, they control a quantity termed marginal FDR which is different from the FDR considered in this paper.

References

Ahmed, A. H. N., Léon, R. and Proschan, F. (1981), Generalized association with applications in multivariate statistics, *Ann. Statist.* **9**, 168 – 176.

Benjamini, Y., Krieger, A. M. and Yekutieli, D. (2006). Adaptive linear step-up procedures that control the false discovery rate. *Biometrika* **93**, 491 – 507.

Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B* **57**, 289–300.

Benjamini, Y. and Yekutieli, D. (2001), The control of the false discovery rate in multiple testing under dependency, *Ann. Statist.* **29**, 1165–1188.

Chi, Z. (2007). On the performance of FDR control: Constraints and a partial solution. *Ann. Statist.* **35**, 1409 – 1431.

Cohen, A. and Sackrowitz, H. B. (2005). Characterization of Bayes procedures for multiple endpoint problems and inadmissibility of the step-up procedure. *Ann. Statist.* **33** 145-158.

Cressie, N. (1979). An optimal statistic based on higher order gaps. *Biometrika* **66**, 619 – 627.

Efron, B. (2004). Selection and estimation for large-scale simultaneous inference. *J. Amer. Statist. Assoc.* **96**, 96 – 104.

Efron, B., Tibshirani, R., Storey, J. D., and Tusher, V. (2001), Empirical Bayes analysis of a microarray experiment, *J. Amer. Statist. Assoc.* **96**, 1151 – 1160.

Ferreira, J. A. and Zwinderman, A. H. (2006). On the Benjamini- Hochberg method. *Ann. Statist.* **34**, 1827 – 1849.

Finner, H., Dickhaus, T. and Roters, M. (2008). On the false discovery rate and an asymptotically optimal rejection curve. *Ann. Statist.* **37**, 596618.

Genovese, C., Wasserman, L. 2002. Operating characteristics and extensions of the false discovery rate procedure. *J. Roy. Statist. Soc. Ser. B* 64, 499 – 517.

Genovese, C. R. and Wasserman, L. (2004). A stochastic process approach to false discovery control. *Ann. Statist.* , 1035 – 1061.

Greenwood, M. (1946). The statistical study of infectious diseases. *J. Roy. Statist. Soc. Ser. A* **109**, 85 – 110.

Jin, J. and Cai, T. (2007). Estimating the null and the proportion of non-null effects in large-scale multiple comparisons. *J. Amer. Statist. Assoc.* **102**, 495-506.

Moran, P. A. P. (1947). The random division of an interval. *J. Roy. Statist. Soc. Ser. B* **9**, 92 – 98.

Nevius, S. E., Proschan, F. and Sethuraman, J. (1977). Schur functions in statistics II. Stochastic majorization. *Ann. Statist.* **5**, 263 – 273.

Pearson, K. (1902). Note on Francis Galton's problem. *Biometrika* **1**, 390 – 399.

Proschan, F. and Sethuraman, J. (1977). Schur functions in statistics I. The Preservation Theorem. *Ann. Statist.* **5**, 256 – 262.

Pyke R. (1965). Spacings (with discussion). *J. Roy. Statist. Soc. Ser. B* **27**, 395449.

Sarkar, S. K. (2002), Some results on false discovery rates in stepwise multiple testing procedures, *Ann. Statist.* **30**, 239 - 257.

Sarkar, S. K. (2006). False discovery and false nondiscovery rates in single-step multiple testing procedures. *Ann. Statist.* **34**, 394 – 415.

Sarkar, S. K. and Guo, W. (2009). On a generalized false discovery rate. *Ann. Statist.* **37**, 1545-1565.

Sarkar, S., Zhou, T., and Ghosh, D. (2008). A general decision-theoretic approach to multiple testing procedures for false discovery and false nondiscovery rates. *Statistica Sinica* **18**, 925 – 946.

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. *J. Roy. Statist. Soc. Ser. B* **66**, 187 – 205.

Table 1. Outcomes of n tests of hypotheses

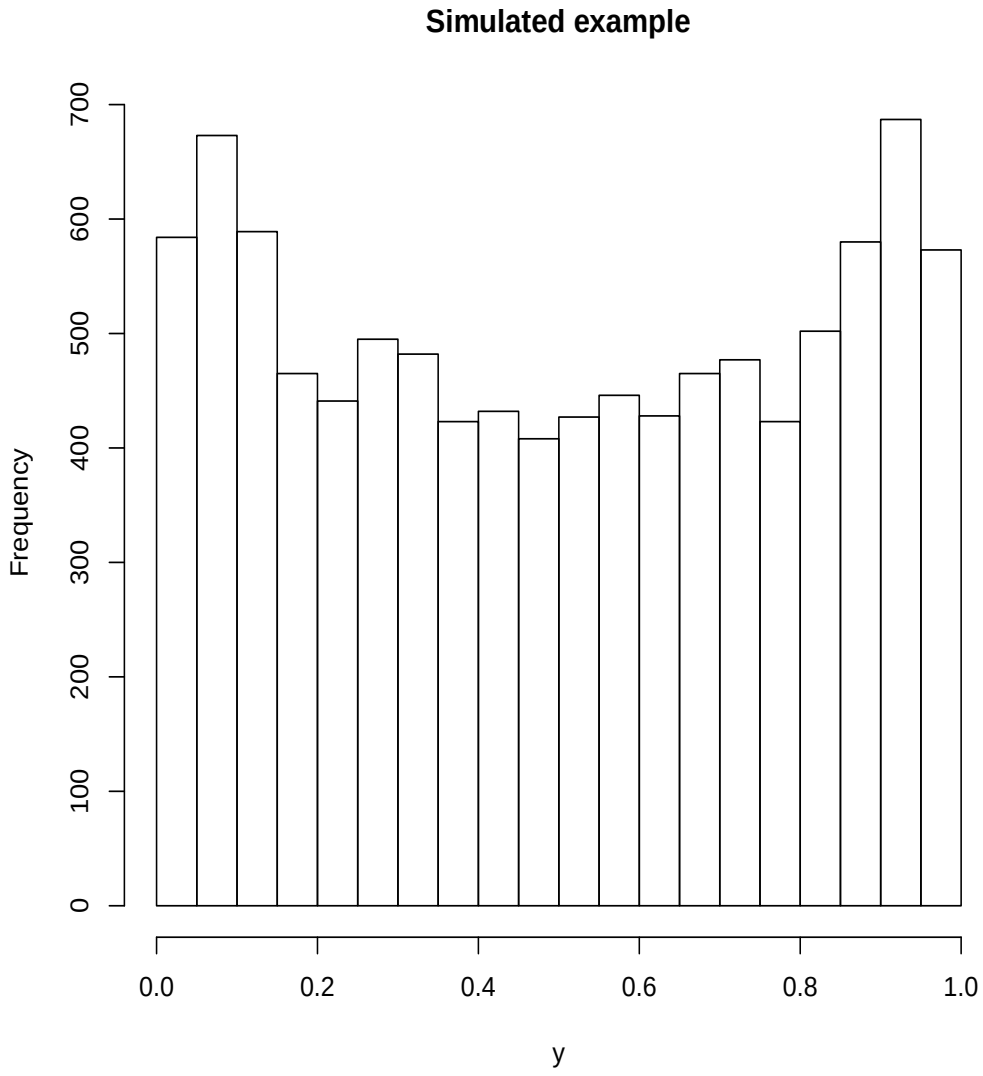
	Accept	Reject	Total
True Null	U	V	n_0
True Alternative	T	S	n_1
	W	Q	n

Table 2. FDR results of simulation studies from Section 6.

	Independent			Dependent		
Method/n_1	20	60	100	20	60	100
BH	0.04	0.04	0.03	0.001	0.005	0.004
Proposed, $\lambda = 2$	0.002	0.03	0.016	0.04	0.025	0.017
Proposed, $\lambda = 4$	0.01	0.04	0.019	0.03	0.06	0.06
Proposed, $\lambda = 16$	0.04	0.04	0.02	0.02	0.08	0.007

Table 3. Power $(1 - NDR)$ results of simulation studies from Section 6.

	Independent			Dependent		
Method/n_1	20	60	100	20	60	100
BH	0.60	0.96	0.99	0.34	0.63	0.71
Proposed, $\lambda = 2$	0.94	1.00	1.00	0.73	0.89	0.93
Proposed, $\lambda = 4$	0.89	0.96	0.99	0.79	0.91	0.94
Proposed, $\lambda = 16$	0.17	0.36	0.49	0.22	0.32	0.47


Simulated example
Frequency
0
100
200
300
400
500
600
700
0.0
0.2
0.4
0.6
0.8
1.0
y