A NEW APPROACH TO FALSE DISCOVERY RATES AND MULTIPLE HYPOTHESIS TESTING

by

John D. Storey

Technical Report No. 2001-18
June 2001

Department of Statistics
STANFORD UNIVERSITY
Stanford, California 94305

A NEW APPROACH TO FALSE DISCOVERY RATES AND MULTIPLE HYPOTHESIS TESTING

by

John D. Storey
Department of Statistics
Stanford University

Technical Report No. 2001-18
June 2001

This research was supported in part by National Institute of Health grant 5R01 CA 72028 and National Science Foundation, Graduate Research Fellowship

Department of Statistics
Sequoia Hall
STANFORD UNIVERSITY
Stanford, California 94305

http://www-stat.stanford.edu

A New Approach to False Discovery Rates and Multiple Hypothesis Testing

John D. Storey†

Department of Statistics, Stanford University, Stanford CA 94305, USA.

Summary. Testing multiple hypotheses involves guarding against much more complicated errors than when testing a single hypothesis. Whereas one typically controls the Type I error rate for a single hypothesis test, the Family Wise Error Rate (FWER) or the False Discovery Rate (FDR) are controlled for multiple hypothesis tests. Therefore, just as in single hypothesis testing, the acceptable error rate is fixed and the rejection region is found to control the error rate. Controlling the FWER or FDR often involves complicated sequential p-value rejection methods based on the observed data. In other words, the rejection region is *estimated* from the data. In this paper we propose the opposite approach – fix the rejection region and then estimate the error rate. This new approach offers increased applicability, accuracy, and power. We apply this methodology to the FDR and provide evidence for its benefits. Also discussed is the calculation of the q-value, which is the FDR analogue of the p-value. Some simple numerical examples are presented that show this new approach can yield over a 10 times increase in power compared to the Benjamini and Hochberg (1995) method. We also briefly discuss how this approach can be applied to other multiple hypothesis testing error measures, such as the FWER.

Keywords: False Discovery Rate, Multiple Hypothesis Testing, q-values, p-values

1. Introduction

The basic paradigm for single hypothesis testing works as follows. We wish to test a null hypothesis H_0 versus an alternative H_1 based on a statistic X. For a given rejection region Γ, we say H_1 is true when $X \in \Gamma$ and H_0 is true when $X \in \Gamma$. A Type I error occurs when $X \in \Gamma$ but H_0 is really true; a Type II error occurs when $X \notin \Gamma$ but H_1 is really true. In order to choose Γ, the acceptable Type I error is set at some level α. Then all rejection regions are considered that have Type I error less than or equal to α. The one that has the lowest Type II error is chosen. Therefore, the rejection region is sought with respect to controlling the *Type I error*. This approach has been fairly successful, and often times one is able to find a rejection region with nearly optimal power (power = 1-Type II error) while maintaining the desired α level Type I error.

When testing multiple hypotheses, the situation becomes much more complicated. Now each test has Type I and Type II errors, and it becomes unclear how we should measure the overall error rate. The first measure to be suggested was the Family Wise Error Rate (FWER). The FWER is the probability of making one or more Type I errors among all the hypotheses. Instead of controlling the probability of a Type I error at level α for each test, the overall FWER is controlled at level α. Nonetheless, α is chosen so that FWER $\leq \alpha$, and then a rejection region Γ is found that maintains level α FWER, but also yields good

†*Address for correspondence:* jstorey@stat.stanford.edu

power. (We will assume for simplicity that each test has the same rejection region, such as would be the case when using the p-values as the statistic.)

In pioneering work, Benjamini and Hochberg (1995) introduced a multiple hypothesis testing error measure called the False Discovery Rate. This quantity is the expected proportion of false positives out of the total number of rejected hypotheses. In many situations, the FWER is much too strict, especially when the number of tests is large. Therefore, the FDR is a more liberal, yet more powerful quantity to control. We will mostly concentrate on the FDR because it shows a lot of promise in modern applications; but for completeness we will motivate our approach in the context of the FWER and FDR.

Suppose there are n hypotheses we wish to test, of which n_0 of the null hypotheses are true. For example, if we reject all p-values $\leq \alpha$, then the FWER is $1 - (1 - \alpha)^{n_0}$. The difficulty of dealing with the FWER arises because we do not know n_0. Therefore, for a chosen α, we have two choices. We can provide *weak control* of the FWER in that we guarantee FWER $\leq \alpha$ when $n = n_0$. *Strong control* of the FWER is provided when FWER $\leq \alpha$ for any n_0. Usually weak control of the FWER is undesirable, so strong control is obtained. In order to find a common rejection region for all tests, each statistic is converted into its p-value. Under the null hypothesis the p-value has a Uniform[0,1] distribution, but it is often the case that the distribution of the p-value is unknown under the alternative distribution.

The situation that has been studied the most is when we reject hypotheses based on the p-values and when each test is independent (Shaffer 1995). Therefore, we will focus on that situation in this paper. The simplest multiple testing procedure that strongly controls the FWER is the Bonferroni method. In this method, the level of each test is set at α/n, and this provides a FWER $\leq \alpha$ for any n_0. There are many methods that have been subsequently developed – see Shaffer (1995) for a review of these methods. Basically, each method was introduced to try to increase the overall power, while strongly maintaining the FWER $\leq \alpha$. These methods usually involve a sequential treatment of the p-values, either starting from the largest and working towards the smallest p-value (called a step-up method), or starting from the smallest p-value and working up (called a step-down method).

Each sequential p-value method tends to be different, and rarely do they offer any immediate insight into exactly what they are accomplishing. This is really what a sequential p-value method tries to do: using the observed data, it estimates the rejection region so that on average the error rate is less than or equal to α. Thus, a sequential p-value method is really an estimation problem. The estimation is made in the FWER case so that the long run frequency that the method yields a false positive is less than or equal to α. In the FDR case, it is made so that the long run frequency of false positives to total number of rejected hypotheses is less than or equal to α.

At the end of a sequential p-value method, we are given an estimate $\hat{k}$ that tells us to reject $p_{(1)}, p_{(2)}, \ldots, p_{(\hat{k})}$, where $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(n)}$ are the ordered, observed p-values. What can we say about $\hat{k}$? Is there any natural way to provide an error measure on this random variable? Usually the answer is "no" because the method to obtain $\hat{k}$ is too complicated. Even if we did have a measure of the variance of $\hat{k}$, what would this really mean? It is a false sense of security in multiple hypothesis testing to think that we have a 100% guaranteed upper bound on the error. The reality is that this process, like any other process in statistics, involves estimation. The more variable the estimate of $\hat{k}$ is, the worse the procedure is going to work in practice. Therefore, the expected value may be that FWER $\leq \alpha$ or FDR $\leq \alpha$, but we do not know how reliable the methods are on a case

by case basis. If point estimation only involved finding unbiased estimators, then the field wouldn't be so successful. Therefore, the reliability of $\hat{k}$ on a case by case basis does matter even though it has not been explored.

Another weakness of multiple hypothesis testing methods is that they try to control the error rate for all values of n_0. Surely there is information about n_0 in the observed p-values. Why not use this information, which will yield a less stringent procedure and more power? Often, the power of the multiple hypothesis testing method decreases with increasing n. This should not be the case, especially when the tests are independent. The larger n, the more information we have about n_0, and this should be used.

In this paper, we propose a new approach to multiple hypothesis testing. We attempt to use more traditional and straightforward statistical ideas to control the error rate of a multiple hypothesis testing procedure. Instead of fixing α and then estimating $\hat{k}$ (i.e., estimating the rejection region), we fix our rejection region and then estimate α. Then a number of rejection regions can be examined and α estimated from each one. A natural objection to this approach is that it does not offer an upper bound on the error rate. Actually, an upper bound is offered in the same sense as the former approach – if we had an infinite number of data sets, then we could get the exact, conservative estimate. Moreover, since in using this new approach we are in the more familiar point estimation situation, we can use the data to estimate n_0, get confidence intervals on our estimate of the error, and gain flexibility in the definition of the error measure.

We will show that our proposed approach is more effective, flexible, and powerful. The multiple hypothesis testing methods we will describe take advantage of more of the information in the data, and they are much simpler to understand. In the next section we more thoroughly discuss the False Discovery Rate (FDR); we also propose a new definition of the FDR. In Section 3 we formulate our approach in the context of the FDR. Section 4 makes a heuristic comparison between the proposed method and that of Benjamini and Hochberg (1995). Section 5 describes several theoretical results pertaining to the proposed approach. Section 6 gives a maximum likelihood estimate interpretation. Section 7 provides numerical results, comparing our method to the old method. Section 8 considers some practical issues of our method. Section 9 describes a quantity called the q-value, which is the FDR analogue of the p-value. Section 10 is the discussion and briefly mentions several possible extensions and applications our approach. Finally, Section 11 provides proofs and technical comments of our results.

2. The False Discovery Rate

As mentioned in the introduction, there are two error measures commonly used in multiple hypothesis testing: the Family Wise Error Rate (FWER) and the False Discovery Rate (FDR). The FWER is the traditional measure used; Benjamini and Hochberg (1995) recently introduced the FDR. The following table summarizes the various outcomes that occur when testing n hypotheses.

	Accept	Reject	Total
Null True	U	V	n_0
Alternative True	T	S	n_1
	$n-R$	R	n

Note that V is the number of Type I errors (or false positives). Therefore, the FWER is defined to be $\mathbf{Pr}(V \geq 1)$. Controlling this quantity offers a very strict error measure. In general, as the number of tests increases, the power decreases when controlling the FWER. Benjamini and Hochberg (1995) defined the FDR to be

$$\mathbf{E}\left[\frac{V}{V+S}\right] = \mathbf{E}\left[\frac{V}{R}\right], \tag{1}$$

that is, the expected proportion of false positives among all rejected hypotheses. Benjamini and Hochberg (1995) and Benjamini and Liu (1999) provide sequential p-value methods to control this quantity. The FDR offers a much less strict multiple testing criterion, and therefore leads to an increase in power.

Large data sets are becoming much more common, where essentially thousands of hypothesis tests have to be performed. Two such examples are wavelet analysis and genomics. In wavelet analysis, one is often faced with the task of finding as many non-zero coefficients as possible, without choosing too many truly zero coefficients. Therefore, the FDR has been useful in developing strategies for choosing many coefficients, among which only a certain fraction are truly zero (Abramovich and Benjamini 1996, and Abramovich et al. 2000). DNA microarrays are one of several new biotechnologies that allow one to perform genome-wide experiments. In particular, microarrays allow the measurement of the expression levels of thousands of genes simultaneously. Tusher et al. (2001) use the FDR as the error controlling method for testing whether over 6000 genes have a difference in expression level between treated and untreated cells. Therefore, the FDR is definitely of interest to the statistics and scientific community.

The definition of the FDR given above is not entirely satisfactory because a problem arises when $R = 0$. This leads to two natural choices for the FDR. Benjamini and Hochberg (1995) chose to use the definition

$$\mathbf{E}\left[\frac{V}{R}\middle| R > 0\right]\mathbf{Pr}(R > 0) \tag{2}$$

because this definition can be controlled by a sequential p-value method. Note however that weak control of the FWER is implicitly embedded in this definition. We will use the following definition

$$\mathbf{E}\left[\frac{V}{R}\middle| R > 0\right]. \tag{3}$$

Benjamini and Hochberg did not use our definition because this quantity is identically 1 when all null hypotheses are true. Recall that the purpose of this paper is to present methodology that fixes the rejection region and estimates the error rate. Therefore, the concern of Benjamini and Hochberg (1995) is not the case here. In Storey (2001), we make an argument for using the latter definition of the FDR. The Benjamini and Hochberg definition can be thought of as "the rate that false discoveries occur", whereas the latter definition can be thought of as "the rate that discoveries are false." Storey (2001) argues that in most situations, one is not interested in cases where no discoveries occur. Thus, we suggest using the second, more appropriate definition. We follow that line of reasoning in this paper and define the FDR as the following.

DEFINITION 1. *We define the False Discovery Rate* – the rate that discoveries are false – *to be:*

$$FDR = \mathbf{E}\left[\frac{V}{R}\middle| R > 0\right]. \tag{4}$$

For notational convenience denote Benjamini and Hochberg's definition of the FDR as FDR_{BH}; FDR will always refer to the new definition.

3. The New Approach Applied to the FDR

In this section, we will show how the proposed approach to multiple hypothesis testing can be applied to the FDR. We first have to present a few simple facts about FDR under independence.

Suppose we are testing n identical hypothesis tests $H_1, H_2, \ldots, H_n$ with independent statistics $X_1, X_2, \ldots, X_n$. We let $H_i = 0$ when null hypothesis i is true, and $H_i = 1$ otherwise. The tests are "identical" in that the $X_i|H_i$ are identically distributed. Therefore, the same rejection region is used for each test. We assume that the H_i are independent Bernoulli random variables with $\mathbf{Pr}(H_i = 0) = \pi_0$ and $\mathbf{Pr}(H_i = 1) = \pi_1$. Also, let Γ be the rejection region for each hypothesis test.

The following theorem is a modification of Theorem 1 from Storey (2001). It allows us to write the FDR in a very simple form that does not depend on n. Note that if we want to assume the number of true null hypotheses n_0 is fixed, we set $\pi_0 = n_0/n$ – Theorem 1 still holds under this modification. Therefore, we will assume without loss of generality that the H_i are random variables as described above.

THEOREM 1. *Suppose n identical simple hypothesis tests are performed with the i.i.d. statistics $X_1, \ldots, X_n$ and rejection region Γ. Also suppose that a null hypothesis is true with* a priori *probability π_0. Then*

$$FDR(\Gamma) = \frac{\pi_0 \cdot \mathbf{Pr}(X \in \Gamma | H = 0)}{\mathbf{Pr}(X \in \Gamma)}, \tag{5}$$

where $\mathbf{Pr}(X \in \Gamma) = \pi_0 \cdot \mathbf{Pr}(X \in \Gamma | H = 0) + \pi_1 \cdot \mathbf{Pr}(X \in \Gamma | H = 1)$.

This paper will be limited to the case where we reject based on independent p-values. See Storey and Tibshirani (2001, in preparation) for a treatment of more general situations. It follows that for p-value based rejections, all rejection regions are of the form $[0, \gamma]$ for some $\gamma \geq 0$. (See Remark 1 in Section 11 for a justification of this.) For the remainder of the paper, instead of denoting rejection regions by the more abstract Γ, we will denote them by γ, which refers to the interval $[0, \gamma]$.

We say the tests are identical in the sense that each test has the same rejection region. Therefore, in terms of p-values we can write the result of Theorem 1 as

$$FDR(\gamma) = \frac{\pi_0 \cdot \mathbf{Pr}(P \leq \gamma | H = 0)}{\mathbf{Pr}(P \leq \gamma)} = \frac{\pi_0 \cdot \gamma}{\mathbf{Pr}(P \leq \gamma)}, \tag{6}$$

where P is the random p-value resulting from any test. Note that under independence the p-values are exchangeable in the sense that each comes from the null distribution (i.e., Uniform[0,1]) with probability π_0 and from the alternative distribution with probability π_1.

It is easiest to think about this in terms of simple versus simple hypothesis tests, but the reasoning can be extended to composite alternative hypotheses. See Remark 2 in Section 11 for more on this.

Since $\pi_0 \cdot n$ of the p-values are expected to be null, then under mild conditions the largest p-values are most likely to come from the null, uniformly distributed p-values. Hence, a good estimate of π_0 is

$$\widehat{\pi}_0 = \frac{\#\{p_i > \beta\}}{(1-\beta)n} \tag{7}$$

for some well chosen β, where $p_1, \ldots, p_n$ are the observed p-values. In this paper, we let $\beta = 1/2$. (For now we assume that β is fixed, however, it makes sense that β can be chosen in some optimal way based on the observed data. This point is discussed more in Section 8.) Efron et al. (2001) use a similar estimate of π_0 in an empirical Bayes method that is directly related to the FDR.

A natural estimate of $\mathbf{Pr}(P \leq \gamma)$ is

$$\widehat{\mathbf{Pr}}(P \leq \gamma) = \frac{\#\{p_i \leq \gamma\}}{n}. \tag{8}$$

Therefore, our overall estimate of the FDR is

$$\widehat{FDR}(\gamma) = \frac{\widehat{\pi_0} \cdot \gamma}{\widehat{\mathbf{Pr}}(P \leq \gamma)} = \frac{\#\{p_i > \beta\} \cdot \gamma}{\#\{p_i \leq \gamma\} \cdot (1-\beta)}. \tag{9}$$

We summarize our proposed approach to multiple hypothesis testing in the context of the FDR below.

Proposed Approach Using the FDR

(a) For the n hypothesis tests, calculate their respective p-values $p_1, \ldots, p_n$.

(b) Fix β at say $\beta = 1/2$ or some other reasonable choice, such as the median of the p-values.

(c) Form the estimates of π_0 and $\mathbf{Pr}(P \leq \gamma)$ as

$$\widehat{\pi}_0 = \frac{\#\{p_i > \beta\}}{(1-\beta)n} \text{ and } \widehat{\mathbf{Pr}}(P \leq \gamma) = \frac{\#\{p_i \leq \gamma\}}{n}.$$

(d) For any rejection region of interest $[0, \gamma]$, the estimated FDR over that region is

$$\widehat{FDR}(\gamma) = \frac{\widehat{\pi_0} \cdot \gamma}{\widehat{\mathbf{Pr}}(P \leq \gamma)} = \frac{\#\{p_i > \beta\} \cdot \gamma}{\#\{p_i \leq \gamma\} \cdot (1-\beta)}.$$

(e) If $\widehat{FDR}(\gamma) > 1$, set $\widehat{FDR}(\gamma) = 1$.

A similar approach can be taken for the FWER, FDR_{BH}, or any other multiple hypothesis testing error measure. It is likely that for a large number of hypothesis tests (where

this approach is most appropriate), one would be interested in the FDR as we defined it. We mention briefly in Section 10 the proposed approach taken for other error measures.

Even though the estimate of the FDR presented in this section is new, the approach has implicitly been taken before. Yekutieli and Benjamini (1999) introduced the idea of estimating the FDR under dependence within the Benjamini and Hochberg (1995) framework. Tusher et al. (2001) take the approach mentioned here – they fix the rejection region and estimate the FDR. In a sense, this paper is a formalization and refinement of their approach to multiple hypothesis testing. Also, Efron et al. (2001) implicitly take this approach and use a very similar estimate within an empirical Bayes framework. One can see from Theorem 1 how our definition of the FDR can be written in a Bayesian form (Storey 2001).

4. A Connection Between the Two Approaches

In this section we present a heuristic connection between the sequential p-value method of Benjamini and Hochberg (1995) and the approach presented in the previous section. The goal is to provide insight into the increased power and effectiveness of our proposed approach.

> The basic point we will make is that using the Benjamini and Hochberg (1995) method to control FDR_{BH} at level α/π_0 is equivalent to (i.e., rejects the same p-values as) using the proposed method to control FDR at level α. The gain in power from our approach is clear – even though $FDR \geq FDR_{BH}$, we control a smaller error rate ($\alpha \leq \alpha/\pi_0$), yet reject the same number of tests.

Let $p_{(1)} \leq \ldots \leq p_{(n)}$ be the ordered, observed p-values for the n hypothesis tests. The method of Benjamini and Hochberg (1995) finds $\widehat{k}$ such that

$$\widehat{k} = \max\{k : p_{(k)} \leq k/n \cdot \alpha\}. \tag{10}$$

Rejecting $p_{(1)}, \ldots, p_{(\widehat{k})}$ provides $FDR_{BH} \leq \alpha$. (See Genovese and Wasserman (2001) for a thorough analysis and interpretation of this sequential p-value method.)

Now suppose we use our method and take the most conservative estimate $\widehat{\pi}_0 = 1$. Then the estimate $\widehat{FDR}$ is less than or equal to α if we reject $p_{(1)}, \ldots, p_{(\widehat{m})}$ such that

$$\widehat{m} = \max\{m : \widehat{FDR}(p_{(m)}) \leq \alpha\}. \tag{11}$$

But since $\widehat{FDR}(p_{(m)}) = \frac{\widehat{\pi}_0 \cdot p_{(m)}}{m/n}$ this equivalent to (with $\widehat{\pi}_0 = 1$)

$$\widehat{m} = \max\{m : p_{(m)} \leq m/n \cdot \alpha\}. \tag{12}$$

Therefore, $\widehat{k} = \widehat{m}$ when $\widehat{\pi}_0 = 1$ even though $FDR_{BH} \leq FDR$. Moreover, if we take the better estimate

$$\widehat{\pi}_0 = \frac{\#\{p_i > \beta\}}{(1-\beta)n} \tag{13}$$

then $\widehat{m} \geq \widehat{k}$, and with high probability $\widehat{m} > \widehat{k}$.

Therefore, we have shown that $\widehat{m} \geq \widehat{k}$ even though $FDR_{BH} \leq FDR$. In other words, using our approach, we reject a greater number of hypotheses while controlling an error measure FDR that is greater than or equal to FDR_{BH}. This leads to greater power while controlling an error rate that is larger yet more appropriate. If we wanted to apply our approach to FDR_{BH}, we would take as our estimate

$$\widehat{FDR}_{BH}(\gamma) = \widehat{FDR}(\gamma) \cdot (1 - (1-\gamma)^n). \tag{14}$$

Note that the term $(1-(1-\gamma)^n)$ is a conservative estimate of $\mathbf{Pr}(R > 0)$ in FDR_{BH}. This clearly leads to an increase in power over the Benjamini and Hochberg (1995) method.

Recall that in calculating FDR the rejection region is fixed. Our comparison is not completely rigorous since it is unknown whether rejecting $p_{(1)}, \ldots, p_{(\widehat{m})}$ provides that $FDR \leq \alpha$. (Numerical evidence that it does, however, is presented in Section 9.) Nevertheless, one can arrive at the Benjamini and Hochberg (1995) algorithm using our results and the most conservative estimate of π_0. Note that the exact same conclusion would be reached if we took the opposite approach and used the Benjamini and Hochberg (1995) algorithm to estimate the FDR, thereby comparing the two approaches.

5. Theoretical Results

In this section, we will provide finite sample and large sample results for $\widehat{FDR}(\gamma)$. Our goal of course is to provide a conservative estimate of $FDR(\gamma)$. In other words we want $\widehat{FDR}(\gamma) \geq FDR(\gamma)$ as much as possible without being too conservative. The following result addresses the finite sample issue of $\mathbf{E}[\widehat{FDR}(\gamma)]$ and its relationship to $FDR(\gamma)$.

THEOREM 2. *For fixed β it follows that*

$$\mathbf{E}[\widehat{FDR}(\gamma)] \geq FDR(\gamma). \tag{15}$$

Therefore, this theorem tells us that the expected value of our estimate is always greater than or equal to the true FDR. This is a comparable result to showing under the former approach that a sequential p-value method controls the FDR. Note that our methodology is especially suited for a large number of hypothesis tests, such as would be encountered in genomics or wavelets. Point estimates from very small samples are quite dangerous anyway, but the same caveat holds for many sequential p-value methods since these methods also involve estimation.

We can also get a large sample result for $\widehat{FDR}(\gamma)$. The tightness to which $\widehat{FDR}(\gamma)$ converges to an upper bound of $FDR(\gamma)$ largely depends on how power changes with Type I error. To this end, let $g(\alpha)$ be the power as a function of Type I error α. For n identical tests, $g(\alpha)$ is the same for each test. If the alternative hypothesis is composite, then $g(\alpha)$ must be defined as the appropriate mixture - see Remark 2 in Section 11. We assume without loss of generality that $g(0) = 0$ and $g(1) = 1$. Also, recall β is the paramter used in the estimation of π_0.

THEOREM 3. *For fixed β we have*

$$\lim_{n \to \infty} \widehat{FDR}(\gamma) = \frac{\pi_0 + \frac{1-g(\beta)}{1-\beta} \cdot \pi_1}{\pi_0} FDR(\gamma) \geq FDR(\gamma) \tag{16}$$

almost surely.

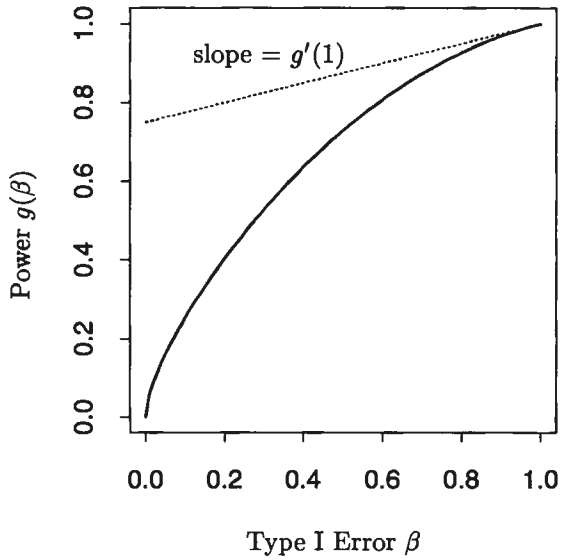

Fig. 1. A plot of power $g(\beta)$ versus Type I error β. It can be seen that since g is concave $\frac{1-g(\beta)}{1-\beta}$ gets smaller as $\beta \to 1$. The line has slope equal to $\lim_{\beta \to 1} \frac{1-g(\beta)}{1-\beta}$, which is the smallest value of $\frac{1-g(\beta)}{1-\beta}$ that can be attained for concave g.

This theorem can be understood graphically in terms of the plot of power to Type I error for each rejection region $[0, \beta]$. The function $g(\beta)$ gives the power over the rejection region $[0, \beta]$, and of course the Type I error over this region is β. The estimate of π_0 is taken over the interval $[\beta, 1]$, so that $1 - g(\beta)$ is the probability of a p-value from the alternative distribution falling into $[\beta, 1]$. Likewise, $1 - \beta$ is the probability of null p-value falling into $[\beta, 1]$. The estimate of π_0 is better the more $g(\beta) > \beta$. This is the case since the interval $[\beta, 1]$ will contain less alternative p-values, and hence the estimate will be less conservative. Figure 1 shows a plot of $g(\beta)$ versus β for a concave g. For concave g, the estimate of π_0 becomes less conservative as $\beta \to 1$. This is formally stated in the following corollary.

COROLLARY 1. *For concave g*

$$\inf_{\beta} \lim_{n \to \infty} \widehat{FDR}(\gamma) \stackrel{a.s.}{=} \lim_{\beta \to 1} \lim_{n \to \infty} \widehat{FDR}(\gamma) \stackrel{a.s.}{=} \frac{\pi_0 + g'(1) \cdot \pi_1}{\pi_0} FDR(\gamma), \tag{17}$$

where $g'(1)$ is the derivative of g evaluated at 1.

In other words, the right hand side of equation (17) is the tightest upper bound $\widehat{FDR}(\gamma)$ can attain on the FDR as $n \to \infty$ for concave g. The corollary can be seen graphically in Figure 2. A plot of $\frac{1-g(\beta)}{1-\beta}$ versus β is shown for a concave g. It can be seen that the minimum is obtained at $\beta = 1$. The minimum value is $g'(1)$, which happens to be $1/4$ in this graph. Whenever the rejection regions are based on a monotone function of the likelihood ratio between the null and alternative hypotheses, g is concave. Note that if g is not concave, then the optimal β used in the estimate of π_0 may not be 1.

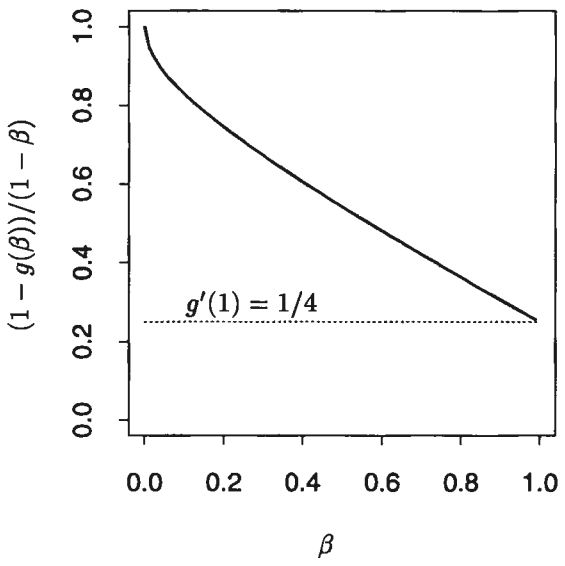

Fig. 2. A plot of $\frac{1-g(\beta)}{1-\beta}$ versus β is shown for a concave g. It can be seen that the minimum is obtained at $\beta = 1$ with value $g'(1) = 1/4$.

A nice property of this last result is that $g'(1) = 0$ whenever testing a single parameter of an exponential family. Therefore, in many of the common cases, we can get exact convergence as $\beta \to 1$.

6. $\widehat{FDR}(\gamma)$ is a Maximum Likelihood Estimate

The approach proposed in this paper involves estimating the FDR. A sensible place to start would be to find the maximum likelihood estimate of $FDR(\gamma)$. Using the notation of the previous section, we have that $\mathbf{Pr}(P \leq \gamma | H = 0) = \gamma$, $\mathbf{Pr}(P \leq \gamma | H = 1) = g(\gamma)$, and $\mathbf{Pr}(H = 0) = \pi_0$. Also, for notational convenience we will let $F(\gamma) = \mathbf{Pr}(P \leq \gamma)$, $X = \#\{p_i : p_i \leq \gamma\}$, and $Y = \#\{p_i : p_i > \beta\}$. We are interested in finding maximum likelihood estimates of π_0 and $F(\gamma)$, so that we may combine them to find a maximum likelihood estimate of $FDR(\gamma)$. It simply follows that the likelihood of the data can be written as

$$F(\gamma)^X \cdot [1 - F(\gamma)]^{n-X} = [\pi_0 \cdot \gamma + (1 - \pi_0) \cdot g(\gamma)]^X [1 - \pi_0 \cdot \gamma - (1 - \pi_0) \cdot g(\gamma)]^{n-X}. \quad (18)$$

Regardless of our knowledge of g, the maximum likelihood estimate of $F(\gamma)$ is

$$\widehat{F}(\gamma) = \frac{X}{n}. \quad (19)$$

If g is known, the maximum likelihood estimate of π_0 is

$$\widetilde{\pi}_0 = \frac{\widehat{F}(\gamma) - g(\gamma)}{\gamma - g(\gamma)} = \frac{X/n - g(\gamma)}{\gamma - g(\gamma)}. \tag{20}$$

Therefore, when g is known, the mle of $FDR(\gamma)$ is

$$\widetilde{FDR}(\gamma) = \frac{\widetilde{\pi}_0 \cdot \gamma}{\widehat{F}(\gamma)} = \frac{[X/n - g(\gamma)]\gamma}{[\gamma - g(\gamma)]\frac{X}{n}}. \tag{21}$$

The behavior of this estimate should be good for large n since it is consistent and efficient.

Recall that in Section 3 we introduced the estimate

$$\widehat{FDR}(\gamma) = \frac{\widehat{\pi_0} \cdot \gamma}{\widehat{F}(\gamma)} = \frac{\frac{Y}{n} \cdot \gamma}{(1-\beta) \cdot \frac{X}{n}}, \tag{22}$$

where $\widehat{\pi}_0 = \frac{Y}{n \cdot (1-\beta)}$ served as our estimate of π_0. This different estimate was used because we did not assume that g was known. Ideally, we would like to find mle's of both π_0 and $g(\gamma)$ when $g(\gamma)$ is unknown. Since $\widehat{F}(\gamma) = X/n$ regardless of the knowledge of g, it follows that the mle of $g(\gamma)$, say $\widetilde{g}(\gamma)$, and $\widetilde{\pi}_0$ would have to satisfy the equation

$$\widetilde{\pi}_0 \cdot \gamma + (1 - \widetilde{\pi}_0) \cdot \widetilde{g}(\gamma) = \frac{X}{n}. \tag{23}$$

This is one equation with two unknowns, so it is impossible to find both mle's simultaneously. Therefore, our remedy was to find a conservative estimate of π_0. Note that when we observe the p-values $p_1, \ldots, p_n$, we can form any reject region $[0, \beta]$. Also note that

$$\frac{1 - F(\beta)}{1 - \beta} = \pi_0 + \frac{1 - g(\beta)}{1 - \beta} \cdot (1 - \pi_0). \tag{24}$$

Without knowing g we can form the mle of $\frac{1-F(\beta)}{1-\beta}$ as Y/n. Therefore, we are estimating a parameter with conservative bias over π_0 of size

$$\frac{1 - g(\beta)}{1 - \beta} \cdot (1 - \pi_0). \tag{25}$$

One could choose $\beta = \gamma$, however this does not have to be the case. Since $\frac{1-g(\beta)}{1-\beta}$ usually gets smaller as β gets larger, it may be better to take a larger β than γ, because γ will likely be very small.

Therefore, the estimate $\widehat{FDR}(\gamma)$ is the maximum likelihood estimate of

$$\frac{\pi_0 + \frac{1-g(\beta)}{1-\beta} \cdot \pi_1}{\pi_0} FDR(\gamma), \tag{26}$$

a quantity slightly greater than $FDR(\gamma)$. In situations where g is unknown, this estimate is, loosely speaking, "optimal" in that the bias can usually be made arbitrarily small (see Corollary 1), while obtaining the smallest asymptotic variance for an estimator of that bias. Moreover, the variance of $\widehat{FDR}(\gamma)$ should not be that different than that of $\widetilde{FDR}(\gamma)$ for large n and powerful tests.

One criticism of the current approach to multiple hypothesis testing that we have made is that the variability of the estimated rejection region resulting from a sequential p-value method is not calculated. Even if it is calculated, it is hard to interpret what $\mathbf{Var}(\widehat{k})$ means in terms of the effectiveness of the sequential p-value method. The only property of $\widehat{k}$ that is assessed is that the expected error rate is less than or equal to α under $\widehat{k}$.

Using our approach, the variance of $\widehat{FDR}(\gamma)$ can easily be calculated. In a parametric situation where g is known, one can calculate $\mathbf{Var}(\widehat{FDR}(\gamma))$ either in the finite sample or asymptotic sense. If g is unknown, the bootstrap can be employed to estimate $\mathbf{Var}(\widehat{FDR}(\gamma))$. Nonetheless, we are guaranteed that as n gets large a minimum variance is attained.

7. A Numerical Comparison

In this section we present some numerical results in order to compare the power of the Benjamini and Hochberg (1995) approach to our proposed approach. As mentioned in the last section, it is not straightforward to compare these two methods since the former estimates the rejection region while the latter estimates the FDR. From Theorem 1, it can be seen, however, that there is a one to one correspondence between the rejection region (parameterized by γ as before) and its corresponding $FDR(\gamma)$. Therefore, we will look at the two values $\gamma = 0.01525, 0.001$ and the corresponding $FDR(\gamma)$ over several values of π_0. The values of γ were chosen in order to cover a wide variety of FDR values.

We will denote the results from the proposed method as "PM", and from the Benjamini and Hochberg method as "BH". For each value of γ and π_0 it is possible to calculate $FDR(\gamma)$. From this we will calculate the resulting average power for BH controlling the FDR at level $FDR(\gamma)$. Likewise, we will calculate the average power resulting from PM over the rejection region $[0, \gamma]$. Recall that the $FDR_{BH} \leq FDR$ so that forcing the BH method to control the FDR at level $FDR(\gamma)$ only helps the BH method.

In the following simulation we performed $n = 1000$ hypothesis tests of $\mu = 0$ versus $\mu = 2$ for i.i.d. random variables $Z_i \sim N(\mu, 1)$, $i = 1, \ldots, 1000$, over 1000 iterations. The null hypothesis for each test is that $\mu = 0$, so the frequency of $Z_i \sim N(0, 1)$ was set to π_0; hence, π_1 of the statistics have the alternative distribution $N(2, 1)$. For each test the p-value is defined as $p_i = \mathbf{Pr}(N(0, 1) \geq z_i)$, where z_i is the observed value of Z_i. In order to calculate the power of PM, test i was rejected if $p_i \leq \gamma$, and the power was calculated accordingly. From the observed p-values, the BH method was performed at level $FDR(\gamma)$, and the rejection region was estimated from the data. The power is calculated from this estimated rejection region. This simulation was performed for $\pi_0 = 0.1, 0.2, \ldots, 0.9$. The corresponding FDR's the BH procedure was used to control are listed in Table 1; these FDR's are calculated using Theorem 1. Two other measurements were also taken – the estimate $\widehat{FDR}(\gamma)$ for PM, and the estimate $\widehat{\gamma}$ for BH.

Even though in this situation we know the alternative distribution of the p-values, we did not use this knowledge. Instead, we estimated the FDR as if the alternative distribution was unknown. (This should also help the BH method in the comparison.) Therefore, we had to choose a value of β in order to estimate π_0; we used $\beta = 1/2$ in all calculations.

Table 1 shows the results of the simulation study. The first half of the table corresponds to $\gamma = 0.01525$, and the second half corresponds to $\gamma = 0.001$. It can be seen that there is a substantial increase in power using the proposed method. One case even gives an 1100% increase in power. The power is constant over each case of PM because the same rejection

Table 1. A numerical comparison between the BH and proposed methods

π_0	FDR	Power (PM)	Power (BH)	$\mathbf{E}[\widehat{FDR}]$ (PM)	$\mathbf{E}[\widehat{FDR}]$ (BH)	$\mathbf{E}[\widehat{\pi}_0]$ (PM)	$\mathbf{E}[\widehat{\gamma}]$ (BH)
$\gamma = 0.01525$							
0.1	0.004	0.435	0.068	0.005	0.0003	0.141	0.0002
0.2	0.008	0.435	0.134	0.010	0.002	0.237	0.0009
0.3	0.015	0.436	0.191	0.016	0.004	0.331	0.002
0.4	0.023	0.435	0.236	0.024	0.009	0.428	0.003
0.5	0.034	0.435	0.277	0.035	0.017	0.521	0.004
0.6	0.050	0.435	0.315	0.052	0.030	0.618	0.006
0.7	0.076	0.435	0.347	0.077	0.053	0.713	0.008
0.8	0.123	0.436	0.377	0.124	0.098	0.807	0.010
0.9	0.240	0.435	0.406	0.243	0.216	0.902	0.012
$\gamma = 0.001$							
0.1	0.0008	0.138	0.011	0.001	8×10^{-6}	0.141	7×10^{-6}
0.2	0.002	0.138	0.026	0.002	0.0004	0.237	3×10^{-5}
0.3	0.003	0.139	0.041	0.003	0.001	0.331	8×10^{-5}
0.4	0.005	0.138	0.056	0.005	0.002	0.428	0.0001
0.5	0.007	0.138	0.071	0.008	0.003	0.521	0.0002
0.6	0.011	0.139	0.087	0.011	0.006	0.618	0.0003
0.7	0.017	0.138	0.101	0.017	0.013	0.713	0.0005
0.8	0.028	0.138	0.116	0.029	0.022	0.807	0.0006
0.9	0.061	0.138	0.129	0.065	0.059	0.902	0.0007

region is used. The power of BH increases as π_0 gets larger because the procedure becomes less conservative. In fact, it can be shown from Section 4 that as $\pi_0 \to 1$, the BH method becomes the PM method.

The fifth column of Table 1 shows $\mathbf{E}[\widehat{FDR}]$ for PM. It can be seen that this is very close to the true FDR (usually within 0.1%), and it is always conservative. The PM method is nearly optimal in that it estimates the $FDR(\gamma)$ basically as close as conservatively possible for each rejection region. Therefore, we essentially lose no power regardless of the value of π_0. Moreover the method gets better as the number of tests increases; the opposite has been true in the past. The seventh column shows $\mathbf{E}[\widehat{\pi}_0]$ for PM. It can be seen that this estimate is always conservative and very close to the actual value.

The sixth column shows $\mathbf{E}[\widehat{FDR}]$ for BH. This was calculated in the following way. The BH method finds $\widehat{k}$ and rejects $p_{(1)}, \ldots, p_{(\widehat{k})}$. This yields an observed false discovery for each iteration. Averaging over all iterations yields $\mathbf{E}[\widehat{FDR}]$ for BH. It can be seen that $\mathbf{E}[\widehat{FDR}] < FDR(\gamma)$ as is expected, however, as π_0 gets smaller, the estimate becomes more and more conservative. This can also be seen in the eighth column in that the average estimate rejection region $\mathbf{E}[\widehat{\gamma}]$ is much smaller than the true γ leading to a decrease in power.

The power comparisons are also shown graphically in Figure 3. The success of this method largely depends on how well we can estimate $FDR(\gamma)$. It is seen in this simulation that the estimates are very good. This is especially due to the fact that the power-Type I error curve is well behaved in the sense discussed in Section 5. Another consideration that must be taken into account is the variance of $\widehat{FDR}(\gamma)$. This is discussed in Section 8 along with several other related topics.

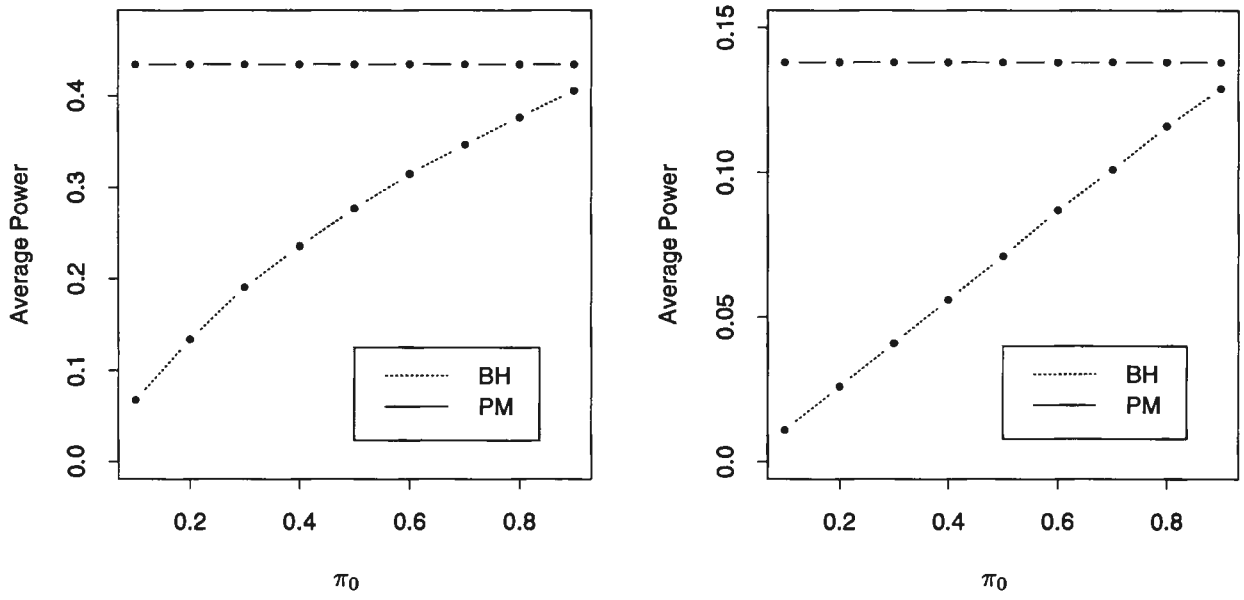

Fig. 3. A plot of average power versus π_0 for the BH method (BH) and the proposed method (PM). The left panel is the case where the rejection region is defined by $\gamma = 0.01525$, and the right panel where $\gamma = 0.001$. It can be seen that there is a substantial increase in power under the proposed method in both situations.

An additional numerical comparison is made in Section 9. In that comparison, we force our estimate into the BH paradigm where the rejection region is estimated. Therefore the comparison is perhaps more direct there. Nevertheless the results are very similar, and the proposed method yields a substantial increase in power.

8. Practical Considerations and Questions

In this section we consider several practical issues of our proposed method. There are many other details that should be investigated in future work, and we describe some of them here.

8.1. Choosing β

As was previously mentioned, a parameter β must be chosen in estimating π_0, the frequency of true null hypotheses. Recall that our estimate is

$$\widehat{\pi}_0 = \frac{\#\{p_i > \beta\}}{(1-\beta)n}. \tag{27}$$

There is an obvious trade-off in the choice of β. When β is small, our estimate is too conservative but very stable. The opposite is true when β is large. Therefore, we want to choose β so that the estimate $\widehat{\pi}_0$ is not too conservative, yet fairly stable.

Throughout this paper, we have used $\beta = 1/2$ simply because it ensures that neither the bias nor the variance of $\widehat{\pi}_0$ is too great. In order to automate this choice in a robust way, we suggest letting β be the median of the observed p-values. If the null p-values happen to be well behaved in the observed data, then taking β to be the median should not be that different than taking $\beta = 1/2$. The case we want to guard against is if all p-values happen to be close to zero in which case $\beta = 1/2$ would be a bad choice. Taking β to be the median guards against that situation.

What we would really like to do is to pick β to minimize some criterion, such as mean squared error. Doing this in a parametric or non-parametric setting will be a topic in our future research (Storey and Tibshirani 2001, in preparation).

8.2. $\widehat{FDR}(\gamma)$ versus $\widetilde{FDR}(\gamma)$

In Section 6, we showed how $\widehat{FDR}(\gamma)$ and $\widetilde{FDR}(\gamma)$ are both maximum likelihood estimates. $\widetilde{FDR}(\gamma)$ is actually the mle of $FDR(\gamma)$, whereas $\widehat{FDR}(\gamma)$ is the mle of a slightly conservative quantity. In order to calculate $\widetilde{FDR}(\gamma)$ one has to know the power function $g(\gamma)$. In terms of mean squared error, where both bias and variance play a role, it is unknown whether $\widetilde{FDR}(\gamma)$ or $\widehat{FDR}(\gamma)$ is better. Note that in most situations $\widehat{FDR}(\gamma)$ should have a smaller variance. It would be interesting to characterize in terms of $g(\gamma)$, when one estimate is better than the other. Also, there may be many other estimates that work well, and hopefully this will be investigated further.

8.3. Calculating the Variance and Confidence Intervals for $\widehat{FDR}(\gamma)$

It is difficult to make general statements about $\mathbf{Var}(\widehat{FDR}(\gamma))$ when g, the power function, is unknown. Here, we will look at the variance from the numerical example in Section

Table 2. The variance of $\widehat{FDR}$

π_0	FDR	$\mathbf{E}[\widehat{FDR}]$	$\mathbf{Var}[\widehat{FDR}]$	π_0	FDR	$\mathbf{E}[\widehat{FDR}]$	$\mathbf{Var}[\widehat{FDR}]$
$\gamma = 0.01525$				$\gamma = 0.001$			
0.1	0.004	0.005	3.41×10^{-7}	0.1	0.0008	0.001	2.21×10^{-8}
0.2	0.008	0.010	7.08×10^{-7}	0.2	0.002	0.002	6.03×10^{-8}
0.3	0.015	0.016	1.40×10^{-6}	0.3	0.003	0.003	1.51×10^{-7}
0.4	0.023	0.024	2.79×10^{-6}	0.4	0.005	0.005	3.50×10^{-7}
0.5	0.034	0.035	6.35×10^{-6}	0.5	0.007	0.008	9.08×10^{-7}
0.6	0.050	0.052	1.40×10^{-5}	0.6	0.011	0.011	2.18×10^{-6}
0.7	0.076	0.077	3.38×10^{-5}	0.7	0.017	0.017	6.89×10^{-6}
0.8	0.123	0.124	1.18×10^{-4}	0.8	0.028	0.029	2.84×10^{-5}
0.9	0.240	0.243	8.26×10^{-4}	0.9	0.061	0.065	3.14×10^{-4}

7. Table 2 lists the 9 observed variances of $\widehat{FDR}(\gamma)$ for $\gamma = 0.01525$ and $\gamma = 0.001$. It can be seen that the variances are of a reasonable size, and provide evidence that accurate estimates of the FDR can be made.

Probably of equal interest to the variance is how to form confidence intervals for the FDR. In a fully parametric situation, realizations of $\widehat{FDR}(\gamma)$ can be simulated and a confidence interval for it can be formed. It can even be analytically calculated based on the maximum likelihood theory presented in Section 6.

The most likely situation is that the alternative distribution is unknown and we must form confidence intervals for $\widehat{FDR}(\gamma)$ non-parametrically. The approach we recommend here is to bootstrap the p-values. For the b^{th} bootstrap sample we form $\widehat{FDR}(\gamma)*b$, $b = 1, \ldots, B$. The desired confidence intervals and standard errors can be formed in the usual way (Efron and Tibshirani 1993).

This technique works well even in situations in which the alternative hypothesis is composite. In the same vein as Remark 2 in Section 11, we can regard the alternative p-values as coming from a mixture distribution over the parameters comprising the alternative hypothesis. This is a natural way to regard the situation since in multiple hypothesis testing we treat the p-values as exchangeable random variables. Therefore, the observed p-values give us the empirical distribution of the mixture of null and alternative p-values, as well as the possible mixture of alternative p-values.

9. The q-value

We now introduce a natural FDR analogue of the p-value, which we call the *q-value*. This quantity was first developed and investigated in Storey (2001). The q-value gives the scientist a hypothesis testing error measure for each observed statistic with respect to the FDR. The p-value accomplishes the same goal with respect to Type I error, and the adjusted p-value with respect to the FWER. We first introduce the q-value in the context of the hypothesis tests performed in Section 7 in the following example.

9.1. Example: Testing the Mean of a $N(\mu, 1)$ Random Variable

Suppose we perform n hypothesis tests of $\mu = 0$ versus $\mu = 2$ for n independent $N(\mu, 1)$ random variables $Z_1, \ldots, Z_n$. Given we observe the random variables to be $Z_1 = z_1, \ldots, Z_n = z_n$, the p-value of $Z_i = z_i$ can be calculated as $p_i = \mathbf{Pr}(Z \geq z_i | H = 0)$. In other words,

it gives the probability of a Type I error if we reject any statistic as extreme or more extreme than z_i. Likewise, if n_0 of the null hypotheses are true, then the adjusted p-value is $\tilde{p}_i = 1 - (1 - p_i)^{n_0}$. This quantity gives the FWER if we reject any statistic as extreme or more extreme than z_i among all n hypotheses.

Now suppose we want to know the FDR if we reject any statistic as extreme or more extreme than z_i among all n hypotheses. By Theorem 1, it follows that it is

$$q_i = \frac{n_0 \mathbf{Pr}(Z \geq z_i | H = 0)}{n_0 \mathbf{Pr}(Z \geq z_i | H = 0) + (n - n_0)\mathbf{Pr}(Z \geq z_i | H = 1)}. \tag{28}$$

It can be seen that q_i is a natural FDR analogue to both the p-value and the adjusted p-value. The relationship between these three quantities $p_i, \tilde{p}_i, q_i$ can also be understood graphically. Figure 4 shows a graph of the $N(0,1)$ and $N(2,1)$ distributions with the point $Z_i = z_i$ marked with a vertical line. The area under the $N(0,1)$ density to the right of the cutoff is p_i. To get the adjusted p-value of z_i, we have to take into account how many null hypotheses there are. Therefore, we use this area and n_0 to get $\tilde{p}_i$ as above. In order to calculate q_i, we need to know both p_i and the area under $N(2,1)$ to the right of the cutoff, which is the power. Thus, we use these two quantities plus n_0 to calculate q_i.

As will be shown below, q_i is what we call the q-value of $Z_i = z_i$. In many situations, it is the FDR obtained when rejecting a statistic as extreme or more extreme than z_i among all n hypotheses. It is helpful to consider this simple example as we formally introduce the q-value.

9.2. Definition of the q-value

Even though we are only considering hypothesis testing with independent p-values, it helps to formally introduce the q-value in a general setting in order to better motivate its definition. We will later define the q-value in terms of p-values. For a nested set of rejection regions $\{\Gamma\}$ (in the previous example, $\{\Gamma\}$ is all sets of the form $[c, \infty)$ for $-\infty \leq c \leq \infty$), the p-value of an observed statistic $X = x$ is defined to be

$$\text{p-value}(x) = \min_{\{\Gamma : x \in \Gamma\}} \mathbf{Pr}(X \in \Gamma | H = 0). \tag{29}$$

This quantity gives a measure of the strength of the observed statistic with respect to making a Type I error – it is the minimum Type I error rate that can occur when rejecting a statistic with value x for the set of nested rejection regions. In a multiple testing situation, one can adjust the p-values of several statistics in order to control the FWER. The adjusted p-values give a measure of the strength of an observed statistic with respect to making one or more Type I error. In an effort to develop a similar concept for the FDR, we make the following definition.

DEFINITION 2. *For an observed statistic $X = x$ define the **q-value** of x to be:*

$$\textit{q-value}(x) = \min_{\{\Gamma :\ x \in \Gamma\}} FDR(\Gamma). \tag{30}$$

In words, the q-value is a measure of the strength of an observed statistic with respect to the FDR – it is the minimum FDR that can occur when rejecting a statistic with value x for the set of nested rejection regions.

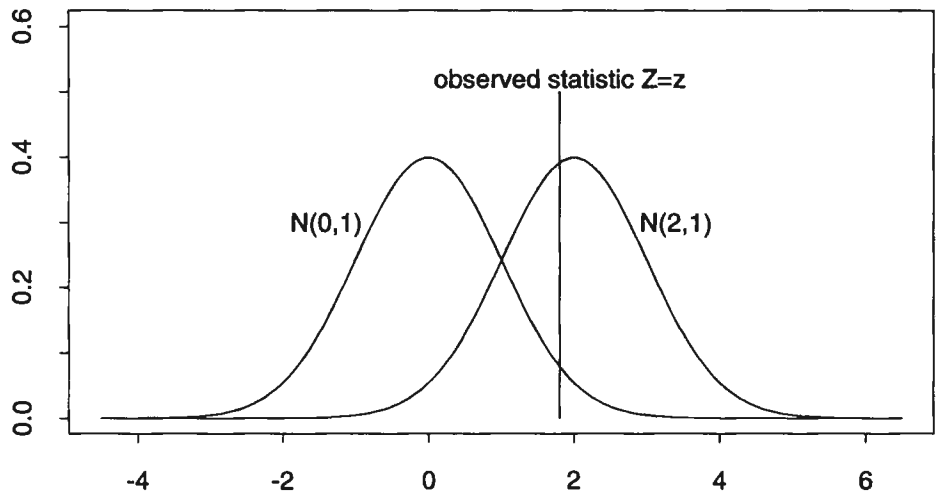

Fig. 4. A plot of the N(0,1) and N(2,1) densities. The vertical line denotes the observed statistic $Z = z$. The p-value and adjusted p-value can be calculated from the area under the N(0,1) density to the right of $Z = z$. The q-value is calculated using the area under both densities to the right of $Z = z$.

The definition is simpler when the statistics are independent p-values. Firstly, the nested set of rejection regions take the form $[0, \gamma]$ and the FDR can be written in a simple form. Therefore in terms of independent p-values, the following is the definition of the q-value of an observed p-value p.

DEFINITION 3. *For a set of hypothesis tests conducted with independent p-values, the q-value of the observed p-value p is*

$$q(p) = \min_{\gamma \geq p} FDR(\gamma) = \min_{\gamma \geq p} \frac{\pi_0 \cdot \gamma}{\mathbf{Pr}(P \leq \gamma)}. \tag{31}$$

A natural property to determine is when it is the case that

$$p = \text{argmin}_{\gamma \geq p} FDR(\gamma), \tag{32}$$

that is, when the most intuitive definition holds that

$$\text{q}(p) = FDR(p) = \frac{\pi_0 \cdot p}{\mathbf{Pr}(P \leq p)}. \tag{33}$$

Storey (2001) answers this question with the following theorem. Recall that we let the function $g(\alpha)$ be the power of each test when using rejection region $[0, \alpha]$.

THEOREM 4. *The q(p) can be written as*

$$q(p) = FDR(p) = \frac{\pi_0 \cdot p}{\mathbf{Pr}(P \leq p)} \tag{34}$$

if and only if g is concave.

9.3. The q-value in practice

According to Theorem 4, we should be able to write $\text{q}(p_i)$ in a very simple form for each p-value obtained in the numerical example considered in this paper since g is concave for likelihood ratio based tests. Note, however, that the result is for the true FDR, and monotonicity does not necessarily hold for the estimated $FDR(p_i)$. Moreover, it will often not be known whether g is a concave function. Therefore, we propose the following algorithm for calculating $\text{q}(p_i)$ in practice. This algorithm comes from the more general definition of the q-value. Recall that for any observed p_i we have $\widehat{FDR}(p_i) = \frac{n\hat{\pi}_0 \cdot p_i}{i}$.

Calculating the q-value

(a) For the n hypothesis tests, calculate the p-values $p_1, \ldots, p_n$.

(b) Let $p_{(1)} \leq \ldots \leq p_{(n)}$ be the ordered p-values.

(c) Set $\text{q}(p_{(n)}) = \widehat{FDR}(p_{(n)})$.

(d) Set $\text{q}(p_{(i)}) = \min\left(\widehat{FDR}(p_{(i)}), \text{q}(p_{(i+1)})\right)$ for $i = n-1, n-2, \ldots, 1$.

Table 3. A comparison between the BH and proposed methods using the q-value

π_0	FDR	Power		$\mathbf{E}[\widehat{FDR}]$		$\mathbf{E}[\widehat{\pi}_0]$	$\mathbf{E}[\widehat{\gamma}]$	
		(PM)	(BH)	(PM)	(BH)	(PM)	(PM)	(BH)
$\gamma = 0.01525$								
0.1	0.004	0.356	0.069	0.003	0.0002	0.141	0.009	0.0002
0.2	0.008	0.398	0.134	0.007	0.002	0.236	0.012	0.0009
0.3	0.015	0.411	0.188	0.013	0.004	0.331	0.013	0.002
0.4	0.023	0.421	0.238	0.021	0.009	0.426	0.014	0.003
0.5	0.034	0.426	0.278	0.032	0.017	0.523	0.014	0.005
0.6	0.050	0.427	0.314	0.049	0.030	0.617	0.014	0.006
0.7	0.076	0.433	0.348	0.074	0.053	0.713	0.015	0.008
0.8	0.123	0.437	0.376	0.122	0.099	0.809	0.015	0.010
0.9	0.240	0.442	0.411	0.238	0.216	0.905	0.015	0.012
$\gamma = 0.001$								
0.1	0.0008	0.102	0.012	0.0004	0.000	0.141	0.0005	7×10^{-6}
0.2	0.002	0.120	0.026	0.001	0.0003	0.236	0.0007	3×10^{-5}
0.3	0.003	0.127	0.041	0.003	0.001	0.331	0.0008	8×10^{-5}
0.4	0.005	0.132	0.057	0.004	0.002	0.426	0.0009	0.0002
0.5	0.007	0.136	0.071	0.007	0.004	0.523	0.0009	0.0002
0.6	0.011	0.136	0.086	0.010	0.006	0.617	0.0009	0.0004
0.7	0.017	0.139	0.101	0.016	0.012	0.713	0.0009	0.0005
0.8	0.028	0.140	0.114	0.028	0.023	0.809	0.0009	0.0006
0.9	0.061	0.149	0.136	0.056	0.050	0.905	0.001	0.0008

This procedure ensures that $\mathrm{q}(p_{(1)}) \leq \ldots \leq \mathrm{q}(p_{(n)})$, which is necessary according to our definition. The way that these q-values can be used in practice is in the following way. It gives us the minimum FDR we can achieve for rejection regions $[0, p_{(i)}]$ for $i = 1, \ldots, n$. In other words, for each p-value there is a rejection region with FDR equal to $\mathrm{q}(p_{(i)})$ so that at least $p_{(1)}, \ldots, p_{(i)}$ are rejected.

9.4. A Numerical Example of the q-value

We will repeat the same numerical example done in Section 7, except here we will let the rejection region be based on the data. In other words, for a given α we will compare the power of the BH method controlling the FDR at level α with the power of the proposed method (PM) rejecting all $\mathrm{q}(p_i) \leq \alpha$. Note that we have not shown any theory that implies rejecting all $\mathrm{q}(p_i) \leq \alpha$ yields $FDR \leq \alpha$, although this is conjectured.

The set-up is the same as in Section 7. We are doing 1000 iterations of $n = 1000$ tests of $\mu = 0$ versus $\mu = 2$ for independent normal random variables $Z_1, Z_2, \ldots, Z_n$. For consistency, we fixed two rejection regions $\gamma = 0.01525, 0.001$ as before, and calculated $FDR(\gamma)$ exactly based on the value of π_0. The two procedure were then applied to control the FDR at level $FDR(\gamma)$. The format of Table 3 is the same as Table 1 in Section 7, except the column $\mathbf{E}[\widehat{FDR}]$ for PM corresponds to the calculated FDR under our current rejection scheme. Note that in all cases the FDR is controlled in the traditional sense under PM. Therefore, this provides evidence that our method also offers control of the FDR within the old paradigm. We have also included a column $\mathbf{E}[\widehat{\gamma}]$ for PM that denotes the average rejection region "estimated" by our proposed method over the 1000 iterations.

It can be seen from Table 3 that we maintain a significant increase in power using

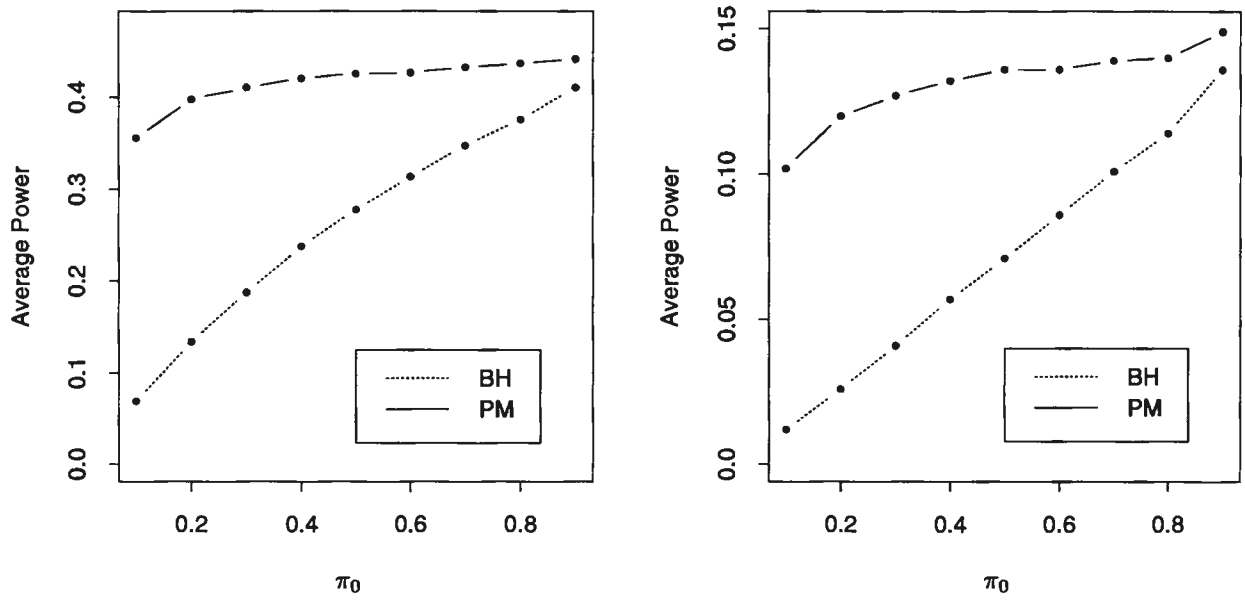

Fig. 5. A plot of average power versus π_0 for the BH method (BH) and the proposed method (PM) forced into the BH paradigm. The left panel is the case where the rejection region is defined by $\gamma = 0.01525$, and the right panel where $\gamma = 0.001$. It can be seen that there is a substantial increase in power under the proposed method in both situations.

our method. The maximum power is almost attained, and we are only slightly below the true rejection region. The results for the BH method are very similar for this simulation because the rejection method is the same. Essentially, we have forced the old paradigm onto our method in this simulation, even though the theory behind our method does not apply towards this. Nonetheless, our method significantly outperforms the BH method. Figure 5 also displays the results of the simulation graphically.

10. Discussion

In this paper, we have proposed a new approach to multiple hypothesis testing. Instead of setting the error rate and estimating the rejection region to control the error at a particular level, we have proposed fixing the rejection region and estimating the error rate. This approach allows a more straightforward analysis of the problem. We have seen that the result is a more powerful and applicable methodology. For example, we proposed a new definition of the FDR, one we feel is usually much more appropriate, and we successfully "controlled" it. By using theoretical results about the FDR with fixed rejection regions, we were able to derive a well behaved estimate of the FDR. Interestingly, the Benjamini and Hochberg (1995) step-up method naturally falls out of these results.

After presenting our ideas about how to address the problem of multiple hypothesis

testing, along with the discussion of the q-value, it is appropriate to reiterate exactly how we recommend applying these ideas. Our methodology has been developed in the context of a fixed rejection region. Therefore, for any rejection region $[0, \gamma]$ we estimate the FDR as

$$\widehat{FDR}(\gamma) = \frac{\widehat{\pi_0} \cdot \gamma}{\widehat{\mathbf{Pr}}(P \leq \gamma)}. \tag{35}$$

Once we observe the p-values $p_1, \ldots, p_n$, it is clear that the rejection regions of interest are $[0, p_i]$ with estimated FDR as $\widehat{FDR}(p_i)$ for $i = 1, \ldots, n$. Also, the q-value for each p-value $\mathrm{q}(p_i)$ can be calculated, which is the FDR analogue of the p-value as explained in the previous section. We have shown in many cases $\mathrm{q}(p_i) = FDR(p_i)$ (Storey 2001), but this does not necessarily occur with observed data. For both quantities, $\widehat{FDR}(p_i)$ and $\mathrm{q}(p_i)$, parametric or non-parametric methods can be applied to get confidence intervals and measures of standard error.

Everything we have discussed in this paper has been under the assumption that we are working with independent p-values. In more general cases, such as with dependence or in non-parametric situations, it is possible to apply very similar ideas to get accurate estimates of the FDR. See Storey and Tibshirani (2001, in preparation) for a forthcoming treatment of this. There are several other open questions that this approach brings to light. For example, a theoretical investigation into $\mathbf{Var}(\widehat{FDR}(\gamma))$ has not been performed. Other, better estimates of the FDR may be available. One could also possibly prove optimality theorems with respect to estimating the FDR within certain frameworks.

We would like to briefly suggest how this approach could be applied to other error rates used in multiple hypothesis testing. For example, given there are n_0 true null hypotheses and the rejection region $[0, \gamma]$, the FWER is $1 - (1 - \gamma)^{n_0}$. This can be conservatively estimated by $1 - (1 - \gamma)^{n \cdot \widehat{\pi}_0}$. Therefore, rejecting all p-values less than or equal to p_i has estimated FWER $1 - (1 - p_i)^{n \cdot \widehat{\pi}_0}$. Using our approach, there is an increase in flexibility. For example, one may wish to use a "robust" measure of multiple testing error, say $\alpha(\pi_0) \cdot FDR(\gamma) + (1 - \alpha(\pi_0)) \cdot FWER(\gamma)$ for some function $\alpha(\pi_0)$. An obvious estimate for this error measure is to take the weighted average of the two suggested here. Most importantly, using our approach really opens the door to a much broader class of multiple testing error measures and procedures.

In a very interesting paper, Friedman (2001) discusses the role statistics can play in the burgeoning field of data mining. Data mining involves investigating huge data sets in which "interesting" features are discovered. The classic example is determining which products tend to be purchased together in a grocery store. It is often the case that the rules for determining interesting features have no simple statistical interpretation. It is understandable that hypothesis testing has not played a major role in this field, because the more hypotheses one has, the less power there is to discover effects. The methodology presented here has the opposite property – the more tests we perform, the better the estimates are. Therefore, it is an asset under this approach to have large datasets with many tests. The only requirement is that the tests have to be exchangeable in the sense that the p-values (or some transformation of the statistics) can be treated homogeneously.

Even if the tests are dependent, our approach can be fully applied. It was shown in Storey (2001) that the effect of dependence is negligible if n is large enough, as long as the dependence is ergodic. Also, Storey and Tibshirani (2001) treat the case where dependence cannot be ignored or where it is infeasible to calculate p-values. Therefore, we hope that

this proposed multiple hypothesis testing paradigm is not only useful in fields like genomics or wavelet analysis, but also in the field of data mining where it is desired to find several interesting features out of many, while limiting the rate of false positives among these.

11. Remarks and Proofs

Remark 1:

Here, we will explain why rejection regions for p-values should be of the form $[0, \gamma]$. Recall that for a nested set of rejection regions $\{\Gamma\}$, the p-value of $X = x$ is defined to be

$$\text{p-value}(x) = \min_{\{\Gamma : x \in \Gamma\}} \mathbf{Pr}(X \in \Gamma | H = 0). \tag{36}$$

Therefore, for two p-values p_1 and p_2, $p_1 \leq p_2$ implies that the respective observed statistics x_1 and x_2 are such that $x_2 \in \Gamma$ implies $x_1 \in \Gamma$. Therefore, whenever p_2 is rejected, p_1 should also be rejected.

Remark 2:

The results presented in this paper are done in the context of simple versus simple hypothesis tests. Suppose instead that the alternative hypothesis is composite. If we view the realized parameters from the alternative distribution as being from a mixture distribution, then the results presented here continue to hold. For example, the probability $\mathbf{Pr}(P \leq \gamma)$ is now not only a mixture of null and alternative p-values, but among the alternative p-values, it is now the appropriate mixture. Also, the power to Type I error function g that was discussed in Section 3 is now also the appropriate mixture of the alternative parameter space. Recall that in multiple hypothesis testing the p-values are treated as exchangeable. Therefore, we also treat the alternative p-values as exchangeable, and it naturally follows to think of them as following a mixture distribution when the alternative hypothesis is composite. See Genovese and Wasserman (2001) for a similar argument.

Proof of Theorem 1:

First note that

$$\begin{aligned} FDR(\Gamma) &= \mathbf{E}\left[\frac{V}{R}\middle| R > 0\right] && (37) \\ &= \sum_{k=1}^{n} \mathbf{E}\left[\frac{V}{R}\middle| R = k\right] \mathbf{Pr}(R = k | R > 0) && (38) \\ &= \sum_{k=1}^{n} \mathbf{E}\left[\frac{V}{k}\middle| R = k\right] \mathbf{Pr}(R = k | R > 0). && (39) \end{aligned}$$

Since the statistics are independent, $V|R = k$ is a binomial random variable with probability of success

$$\frac{\pi_0 \cdot \mathbf{Pr}(X \in \Gamma | H = 0)}{\mathbf{Pr}(X \in \Gamma)}. \tag{40}$$

Therefore,

$$FDR(\Gamma) = \sum_{k=1}^{n} \frac{k \cdot \frac{\pi_0 \cdot \mathbf{Pr}(X \in \Gamma | H=0)}{\mathbf{Pr}(X \in \Gamma)}}{k} \mathbf{Pr}(R = k | R > 0) \tag{41}$$

$$= \frac{\pi_0 \cdot \mathbf{Pr}(X \in \Gamma | H = 0)}{\mathbf{Pr}(X \in \Gamma)}. \tag{42}$$

Proof of Theorem 2:

We will first prove the result conditional on n_0 and n_1, the number of null and alternative hypotheses. Let $X = \#\{\text{null } p_i : p_i > \beta\}$, $Y = \#\{\text{null } p_i : p_i \leq \gamma\}$, and $Z = \#\{\text{alternative } p_i : p_i \leq \gamma\}$. Then

$$\widehat{FDR}(\gamma) \geq \frac{X/(1-\beta) \cdot \gamma}{Y + Z}. \tag{43}$$

Conditioning on X, and using Jensen's inequality on $Y + Z$ we get

$$\mathbf{E}\left[\frac{X/(1-\beta) \cdot \gamma}{Y + Z} \middle| X \right] \geq \frac{X/(1-\beta) \cdot \gamma}{\mathbf{E}(Y|X) + \mathbf{E}(Z)}. \tag{44}$$

Since $\mathbf{E}(Y|X) = \frac{\gamma}{\beta}(n_0 - X)$, we get

$$\frac{X/(1-\beta) \cdot \gamma}{\mathbf{E}(Y|X) + \mathbf{E}(Z)} = \frac{X/(1-\beta) \cdot \gamma}{\frac{\gamma}{\beta}(n_0 - X) + \mathbf{E}(Z)} \tag{45}$$

Using Jensen's inequality on X implies

$$\mathbf{E}\left[\frac{X/(1-\beta) \cdot \gamma}{Y + Z} \right] \geq \frac{\mathbf{E}(X)/(1-\beta) \cdot \gamma}{\frac{\gamma}{\beta}(n_0 - \mathbf{E}(X)) + \mathbf{E}(Z)} \tag{46}$$

Therefore,

$$\mathbf{E}[\widehat{FDR}(\gamma)] \geq \frac{\mathbf{E}(X)/(1-\beta) \cdot \gamma}{\frac{\gamma}{\beta}(n_0 - \mathbf{E}(X)) + \mathbf{E}(Z)} = \frac{\pi_0 \cdot \gamma}{\mathbf{Pr}(P \leq \gamma)} = FDR(\gamma). \tag{47}$$

The result follows unconditionally on n_0 by replacing n_0 with the Binomial(n, π_0) random variable N_0 and applying Jensen's inequality one more time.

Proof of Theorem 3:

Recall that

$$\widehat{FDR}(\gamma) = \frac{\widehat{\pi_0} \cdot \gamma}{\widehat{\mathbf{Pr}}(P \leq \gamma)} = \frac{\#\{p_i > \beta\}/n \cdot \gamma}{\#\{p_i \leq \gamma\}/n \cdot (1-\beta)}. \tag{48}$$

By the Strong Law of Large Numbers, $\widehat{\mathbf{Pr}}(P \leq \gamma) \to \mathbf{Pr}(P \leq \gamma)$ almost surely. Also, $\mathbf{Pr}(P \geq \beta | H = 0) = 1 - \beta$ and $\mathbf{Pr}(P \geq \beta | H = 1) = 1 - g(\beta)$, where $g(\beta)$ is the power of rejecting over $[0, \beta]$ as described in Section 3. Therefore, by the Strong Law of Large Numbers $\#\{p_i \geq \beta\}/n \to (1-\beta) \cdot \pi_0 + (1 - g(\beta)) \cdot \pi_1$ almost surely. Thus, it follows that

$$\lim_{n \to \infty} \widehat{FDR}(\gamma) = \frac{(\pi_0 + \frac{1-g(\beta)}{1-\beta} \cdot \pi_1) \cdot \gamma}{\mathbf{Pr}(P \leq \gamma)} = \frac{\pi_0 + \frac{1-g(\beta)}{1-\beta} \cdot \pi_1}{\pi_0} FDR(\gamma) \geq FDR(\gamma). \tag{49}$$

Proof of Corollary 1:

Since $g(\beta)$ is concave in β, $\frac{1-g(\beta)}{1-\beta}$ is non-increasing in β. Therefore, the minimum of $\frac{1-g(\beta)}{1-\beta}$ is obtained at $\lim_{\beta \to 1} \frac{1-g(\beta)}{1-\beta}$. By L'Hopital's rule, $\lim_{\beta \to 1} \frac{1-g(\beta)}{1-\beta} = g'(1)$.

Acknowledgments

Thanks to Bradley Efron and Ji Zhu for helpful comments and suggestions. I am especially grateful for the ideas and encouragement of my advisor, Robert Tibshirani. This research was supported in part by a NSF Graduate Research Fellowship.

References

Abramovich, F. and Benjamini, Y. (1996) Adaptive thresholding of wavelet coefficients. *Computational Statistics and Data Analysis* **22:** 351-361.

Abramovich, F., Benjamini, Y., Donoho, D., and Johnstone, I. (2000) Adapting to unknown sparsity by controlling the false discovery rate. Technical Report 2000-19, Department of Statistics, Stanford University.

Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B* **57:** 289-300.

Benjamini, Y. and Liu, W. (1999) A step-down multiple hypothesis procedure that controls the false discovery rate under independence. *Journal of Statistical Planning and Inference* **82:** 163-170.

Efron, B. and Tibshirani, R.J. (1993) *An Introduction to the Bootstrap.* Chapman & Hall: New York.

Efron, B., Tibshirani, R., Storey, J.D., and Tusher, V. (2001) Empirical Bayes analysis of a microarray experiment. Technical Report No. 216, Division of Biostatistics, Stanford University.

Friedman, J.H. (2001) The role of statistics in the data revolution? *International Statistical Review*, **69:** 5-10.

Genovese, C. and Wasserman, L. (2001) Operating characteristics and extensions of the FDR procedure. Technical Report, Department of Statistics, Carnegie Mellon University.

Shaffer, J. (1995) Multiple hypothesis testing. *Annual Review of Psychology*, **46:** 561-584.

Storey, J.D. (2001) The False Discovery Rate: A Bayesian interpretation and the q-value. Technical Report No. 2001-12, Department of Statistics, Stanford University.

Storey, J.D. and Tibshirani, R.J. (2001) Estimating the false discovery rate, in preparation.

Tusher, V., Tibshirani, R., and Chu, G. (2001) Significance analysis of microarrays applied to transcriptional responses to ionizing radiation. *Proceedings of the National Academy of Sciences* **98:** 5116-5121.

Westfall, P.H. and Young, S.S. (1993) *Resampling-based multiple testing: examples and methods for p-value adjustment.* Wiley Series in Probability and Mathematical Statistics. Wiley: New York.

Yekutieli, D. and Benjamini, Y. (1999) Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. *Journal of Statistical Planning and Inference* **82:** 171-196.