The Annals of Statistics
2003, Vol. 31, No. 6, 2013–2035

THE POSITIVE FALSE DISCOVERY RATE: A BAYESIAN INTERPRETATION AND THE q-VALUE[1]

BY JOHN D. STOREY

University of Washington

Multiple hypothesis testing is concerned with controlling the rate of false positives when testing several hypotheses simultaneously. One multiple hypothesis testing error measure is the false discovery rate (FDR), which is loosely defined to be the expected proportion of false positives among all significant hypotheses. The FDR is especially appropriate for exploratory analyses in which one is interested in finding several significant results among many tests. In this work, we introduce a modified version of the FDR called the "positive false discovery rate" (pFDR). We discuss the advantages and disadvantages of the pFDR and investigate its statistical properties. When assuming the test statistics follow a mixture distribution, we show that the pFDR can be written as a Bayesian posterior probability and can be connected to classification theory. These properties remain asymptotically true under fairly general conditions, even under certain forms of dependence. Also, a new quantity called the "q-value" is introduced and investigated, which is a natural "Bayesian posterior p-value," or rather the pFDR analogue of the p-value.

1. Introduction. When testing a single hypothesis, one is usually concerned with controlling the false positive rate while maximizing the probability of detecting an effect when one really exists. In statistical terms, we maximize the power conditional on the Type I error rate being at or below some level. The field of multiple hypothesis testing tries to extend this basic paradigm to the situation where several hypotheses are tested simultaneously. One must define an appropriate compound error measure according to the rate of false positives one is willing to encounter. Then a procedure is developed that allows one to control the error rate at a desired level, while maintaining the power of each test as much as possible.

The most commonly controlled quantity when testing multiple hypotheses is the family wise error rate (FWER), which is the probability of yielding one or more false positives out of all hypotheses tested. The most familiar example of this is the Bonferroni method. If there are m hypothesis tests, each test is controlled so that the probability of a false positive is less than or equal to α/m for some chosen value of α. It then follows that the overall FWER is less than or equal to α. Many

Received May 2001; revised March 2003.

[1]Supported in part by an NSF Graduate Research Fellowship.

AMS 2000 subject classification. 62F03.

Key words and phrases. Multiple comparisons, pFDR, pFNR, p-values, q-values, simultaneous inference.

more methods have been introduced that improve upon the Bonferroni method in that the FWER is still controlled at level α, but the average power among the tests is increased. Shaffer (1995) provides a review of many of these methods.

The FWER offers an extremely strict criterion which is not always appropriate. It is possible for a multiple hypothesis testing situation to exist in which one is more concerned about the rate of false positives among all rejected hypotheses rather than the probability of making one or more Type I errors. We have seen a recent increase in the size of data sets available. It is now often up to the statistician to find as many interesting features in a data set as possible rather than test a very specific hypothesis on one item. For example, one is more frequently faced with the daunting task of estimating or performing hypothesis tests on thousands of parameters simultaneously. In this kind of situation, one is more interested in the total number of false positives compared to the total number of significant items, rather than making one or more Type I errors.

Consider Table 1 giving the various outcomes that occur when m hypothesis tests are performed according to some significance rule, which can either be fixed or data-dependent. The FWER can formally be written as $\Pr(V \geq 1)$. In a seminal paper, Benjamini and Hochberg (1995) introduce a new multiple hypothesis testing error measure called the false discovery rate (FDR), which they define as

$$\begin{aligned} \mathrm{FDR} &= \mathrm{E}\left[\frac{V}{R \vee 1}\right] \\ &= \mathrm{E}\left[\frac{V}{R}\middle| R > 0\right] \Pr(R > 0). \end{aligned} \tag{1.1}$$

The only effect of the "$R \vee 1$" in the denominator of the first expectation is that the ratio V/R is set to zero when $R = 0$. Benjamini and Hochberg (1995) prove that a particular p-value step-up method strongly controls the FDR when the true null hypotheses are simple and independent, with an extension to "positive regression dependence" in Benjamini and Yekutieli (2001). This procedure was originally introduced by Simes (1986) to weakly control the FWER. When using this procedure, the realized V and R depend on the random outcome of a p-value-based algorithm.

The Benjamini and Hochberg (1995) procedure works as follows. Suppose that

TABLE 1
Possible outcomes from m hypothesis tests

	Accept null	**Reject null**	**Total**
Null true	U	V	m_0
Alternative true	T	S	m_1
	W	R	m

the p-values resulting from the m tests are ordered such that $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$. If we calculate

$$\widehat{k} = \arg\max_{1 \leq k \leq m} \{k : p_{(k)} \leq \alpha \cdot k/m\},$$

then rejecting the null hypotheses corresponding to $p_{(1)}, \ldots, p_{(\widehat{k})}$ provides $\mathrm{FDR} = m_0/m \cdot \alpha \leq \alpha$. If no p-value satisfies this inequality, then no hypothesis test is called significant. (It is important to keep in mind that for any given set of data we do not have $V/R \leq \alpha$. Rather, the long-run behavior of this procedure is such that $\mathrm{FDR} \leq \alpha$.) The FDR offers less stringent control over Type I errors than the FWER, and is therefore usually more powerful, as is shown in their simulations.

In this paper, we define the *positive false discovery rate* (pFDR) to be $\mathrm{pFDR} = \mathrm{E}[V/R|R > 0]$. The term "positive" describes the fact that we have conditioned on at least one positive finding having occurred. See Section 2 for the motivation and definition. The aim of this paper is to investigate the statistical properties of the pFDR. The following are the main results of this paper.

1. When assuming that the test statistics come from a random mixture of the null and alternative distributions, the pFDR can be written as a simple Bayesian posterior probability (Section 3).
2. The pFDR can be used to define the q-value, which is a natural pFDR (or Bayesian) analogue to the p-value (Section 4).
3. Under fairly general conditions, even certain forms of dependence, the realized V/R, the FDR, and the pFDR converge simultaneously over all significance regions to the Bayesian form of the pFDR (Section 5).
4. The pFDR has a connection to classification theory, and the set of Bayes rules can be used to minimize $(1 - w) \cdot \mathrm{pFDR} + w \cdot \mathrm{pFNR}$, where the pFNR (defined later) is the natural counterpart to the pFDR (Section 6).

Benjamini and Hochberg (1995) define the FDR and prove by an induction argument that the Simes procedure [Simes (1986)] strongly controls the FDR. The goal of this paper is to elucidate false discovery rate quantities, rather than provide estimation techniques, by investigating the pFDR and making connections to other ideas in statistics. For example, hypothesis testing is traditionally known as a frequentist procedure. However, classical classification theory seems to be a bridge between Bayesian modeling and hypothesis testing. In the context of the pFDR, this bridge becomes clearer. The pFDR offers the potential to be a tool for simultaneous decision making useful to both frequentists and Bayesians.

2. The positive false discovery rate. Benjamini and Hochberg (1995) define the FDR according to equation (1.1). The most obvious definition of a false discovery rate is $\mathrm{E}[V/R]$. However, in most cases there is positive probability that $R = 0$, so this definition is not well-defined. Three quantities that remedy the

$R = 0$ problem are

$$\text{(A)} \qquad \mathrm{E}\left[\frac{V}{R}\middle| R > 0\right] \cdot \Pr(R > 0),$$

$$\text{(B)} \qquad \mathrm{E}\left[\frac{V}{R}\middle| R > 0\right],$$

$$\text{(C)} \qquad \frac{\mathrm{E}[V]}{\mathrm{E}[R]}.$$

Benjamini and Hochberg (1995) briefly consider all these quantities and discuss the reasons for their choice of (A). They point out that if all null hypotheses are true, $m_0 = m$, then $\mathrm{E}[V/R|R > 0] = 1$ and $\mathrm{E}[V]/\mathrm{E}[R] = 1$, so neither quantity can be controlled in the traditional p-value-based framework [i.e., when $m_0 = m$ then (B) = (C) = 1. Therefore, one can never choose a value $\alpha < 1$ and guarantee that regardless of m_0, (B) and (C) are less than or equal to α]. Therefore, they choose to work with definition (A). Definition (C) has the advantage of being simple, but the other two quantities measure the joint behavior of V and R.

We argue that when $m_0 = m$, one would want the false discovery rate to be 1, and that one is not interested in cases where no test is significant. Shaffer (1995) also believes that the inclusion of $\Pr(R > 0)$ in the definition of FDR is unsatisfying. These considerations lead us to propose definition (B) as an alternative definition, called the *positive false discovery rate* (pFDR). The modified definition is intuitively pleasing and is shown to be mathematically tractable. We call (B) the pFDR because it is conditioned on the fact that at least one positive finding has occurred.

DEFINITION 1. The *positive false discovery rate* is defined to be

$$\mathrm{pFDR} = \mathrm{E}\left[\frac{V}{R}\middle| R > 0\right].$$

There are two clear approaches to false discovery rates that can be taken. The first is to fix the acceptable rate α beforehand and estimate a significance threshold to obtain this rate conservatively on average. The second is to fix the significance threshold and provide a conservative estimate of the rate over that threshold. When taking the first approach, one is forced to use the FDR since the pFDR cannot be controlled in this sense. The pFDR can be conservatively estimated in the second approach, however. By considering false discovery rates for fixed significance regions, one can gain insight into the operating characteristics of the quantities, resulting in improved procedures. Using the results of an earlier version of this paper [Storey (2001)], Storey (2002a) develops conservative point estimates for both the FDR and pFDR for a fixed significance threshold.

Storey's (2002a) method shows improvements in power over the Benjamini and Hochberg (1995) methodology, mainly due to the estimation of m_0/m. These

estimates have also been shown rigorously to be conservative from a variety of viewpoints [Storey (2002a) and Storey, Taylor and Siegmund (2004)]. Storey, Taylor and Siegmund (2004) show that the point estimates can easily be translated into an FDR controlling method, and they conservatively estimate the FDR and pFDR over all significance regions *simultaneously*. Clearly, the simultaneous conservative estimation is the most useful, as it allows the researcher to perform a truly exploratory analysis. The adjustments of Benjamini and Hochberg (2000), which are similar in spirit to Storey (2002a), have not been shown to provide strong control. Genovese and Wasserman (2002b) model pFDR as a stochastic process over all possible thresholds. From these recent advances, it is clear that the pFDR is a useful quantity to study, and one can overcome some of Benjamini and Hochberg (1995) concerns.

REMARK. An example where confusion in the interpretation of FDR and pFDR is dangerous is the following. One can use the Benjamini and Hochberg (1995) procedure to yield on average that $\mathrm{FDR} \leq 0.1$. But if $\Pr(R > 0) = 0.5$, then we have actually only controlled $\mathrm{pFDR} \leq 0.2$, a quantity twice as large. One may suppose this example is hypothetical, but this exact confusion arises in Weller, Song, Heyen, Lewin and Ron (1998). Zaykin, Young and Westfall (2000) point out that the results of Weller, Song, Heyen, Lewin and Ron (1998) can be very misleading if FDR and pFDR are confused.

3. A Bayesian interpretation. In this section we present a Bayesian interpretation of the pFDR. As it turns out, the pFDR can be written as a simple posterior probability under certain assumptions. Note that throughout this work the pFDR is calculated over a fixed significance region, rather than a data-dependent threshold. Suppose we wish to perform m identical tests of a null hypothesis versus an alternative hypothesis based on the statistics $T_1, T_2, \ldots, T_m$. For a given significance region Γ, define the positive false discovery rate as we defined it in Section 2:

$$\mathrm{pFDR}(\Gamma) = \mathrm{E}\left[\frac{V(\Gamma)}{R(\Gamma)} \middle| R(\Gamma) > 0\right], \tag{3.1}$$

where $V(\Gamma) = \#\{\text{null } T_i : T_i \in \Gamma\}$ and $R(\Gamma) = \#\{T_i : T_i \in \Gamma\}$. Let $H_i = 0$ when the ith null hypothesis is true and $H_i = 1$ when it is false, $i = 1, \ldots, m$. Let π_0 be the a priori probability that a hypothesis is true: that is, we assume that the H_i are i.i.d. Bernoulli random variables with $\Pr(H_i = 0) = \pi_0$ and $\Pr(H_i = 1) = 1 - \pi_0 =: \pi_1$.

Before we present the Bayesian form of the pFDR, consider the pFDR when $m = 1$. Under the above assumptions, the probability of a false positive given that the statistics are significant is $\Pr(H = 0 | T \in \Gamma)$. Also, given that $T \in \Gamma$, $V(\Gamma)/R(\Gamma)$ is 0 or 1 according to whether it is a true positive or false positive, respectively. Therefore, it is easily seen that $\mathrm{pFDR}(\Gamma) = \Pr(H = 0 | T \in \Gamma)$ when $m = 1$. We now show that this result does not change when $m > 1$.

THEOREM 1. *Suppose m identical hypothesis tests are performed with the statistics $T_1, \ldots, T_m$ and significance region Γ. Assume that (T_i, H_i) are i.i.d. random variables, $T_i|H_i \sim (1 - H_i) \cdot F_0 + H_i \cdot F_1$ for some null distribution F_0 and alternative distribution F_1, and $H_i \sim \text{Bernoulli}(\pi_1)$ for $i = 1, \ldots, m$. Then*

$$\text{pFDR}(\Gamma) = \Pr(H = 0|T \in \Gamma), \tag{3.2}$$

where $\pi_0 = 1 - \pi_1$ is the implicit prior probability used in the above posterior probability.

It is surprising that the pFDR, a compound error measure, can be written in such a simple way. Moreover, the posterior probability (3.2) does not depend on m. [Also note that $\Pr(H_i = 0|T_i \in \Gamma)$ is the same for each $i = 1, \ldots, m$, which is why we left out the index in the statement of the theorem.] We can write explicitly

$$\begin{aligned}\text{pFDR}(\Gamma) &= \Pr(H = 0|T \in \Gamma) \\ &= \frac{\pi_0 \cdot \Pr(T \in \Gamma|H = 0)}{\pi_0 \cdot \Pr(T \in \Gamma|H = 0) + \pi_1 \cdot \Pr(T \in \Gamma|H = 1)} \\ &= \frac{\pi_0 \cdot \{\text{Type I error of } \Gamma\}}{\pi_0 \cdot \{\text{Type I error of } \Gamma\} + \pi_1 \cdot \{\text{Power of } \Gamma\}}.\end{aligned}$$

This shows that the pFDR increases with increasing Type I errors and decreases with increasing power. We now prove Theorem 1.

PROOF OF THEOREM 1. First note that

$$\begin{aligned}\text{pFDR}(\Gamma) &= \text{E}\left[\frac{V(\Gamma)}{R(\Gamma)}\middle| R(\Gamma) > 0\right] \\ &= \sum_{k=1}^{m} \text{E}\left[\frac{V(\Gamma)}{R(\Gamma)}\middle| R(\Gamma) = k\right] \Pr\big(R(\Gamma) = k|R(\Gamma) > 0\big) \\ &= \sum_{k=1}^{m} \text{E}\left[\frac{V(\Gamma)}{k}\middle| R(\Gamma) = k\right] \Pr\big(R(\Gamma) = k|R(\Gamma) > 0\big).\end{aligned}$$

Since the statistics are independent, it intuitively follows that $V(\Gamma)|R(\Gamma) = k$ is a binomial random variable with probability of success $\Pr(H = 0|T \in \Gamma)$, in which case the proof easily follows. However, we can be more precise. Because of the i.i.d. assumption, it follows that

$$\begin{aligned}\text{E}[V(\Gamma)|R(\Gamma) = k] &= \text{E}\left[\sum_{i=1}^{m} \mathbb{1}(T_i \in \Gamma)\mathbb{1}(H_i = 0)\middle| \begin{matrix} T_1, \ldots, T_k \in \Gamma \\ T_{k+1}, \ldots, T_m \notin \Gamma \end{matrix}\right] \\ &= \text{E}\left[\sum_{i=1}^{k} \mathbb{1}(H_i = 0)\middle| \begin{matrix} T_1, \ldots, T_k \in \Gamma \\ T_{k+1}, \ldots, T_m \notin \Gamma \end{matrix}\right]\end{aligned}$$

$$= \sum_{i=1}^{k} \mathrm{E}[\mathbb{1}(H_i = 0) | T_i \in \Gamma]$$
$$= k \cdot \Pr(H = 0 | T \in \Gamma).$$

Therefore,

$$\mathrm{pFDR}(\Gamma) = \sum_{k=1}^{m} \frac{k \cdot \Pr(H = 0 | T \in \Gamma)}{k} \Pr\big(R(\Gamma) = k | R(\Gamma) > 0\big)$$
$$= \Pr(H = 0 | T \in \Gamma). \qquad \square$$

Note that when the H_i are not random, this theorem no longer holds, since there is the deterministic constraint that $\sum_{i=1}^{m} H_i = m_1$. However, for large m, a result analogous to Theorem 1 holds under certain convergence assumptions; this is formally dealt with in Section 5. Also, this theorem holds for a simple versus simple test, but composite hypotheses can also be considered as long as one models the alternative parameter as a random variable. Then F_1 is simply the mixture of the alternative distributions.

Recall that in Section 2 we considered three definitions, the third of which was $\mathrm{E}[V]/\mathrm{E}[R]$. The following shows that this quantity is equal to the pFDR under the assumptions of Theorem 1. The proof follows straightforwardly by noting that $\mathrm{E}[V(\Gamma)] = m \cdot \pi_0 \cdot \Pr(T \in \Gamma | H = 0)$ and $\mathrm{E}[R(\Gamma)] = m \cdot \Pr(T \in \Gamma)$.

COROLLARY 1. *Under the assumptions of Theorem* 1,

$$\mathrm{E}\left[\frac{V(\Gamma)}{R(\Gamma)} \Big| R(\Gamma) > 0\right] = \frac{\mathrm{E}[V(\Gamma)]}{\mathrm{E}[R(\Gamma)]}.$$

Even with the existence of this result, we reiterate that in general pFDR captures the joint behavior of V and R, whereas $\mathrm{E}[V]/\mathrm{E}[R]$ does not.

The pFDR written as $P(H = 0 | T \in \Gamma)$ can be related to the Type I error. One could call it a "posterior Bayesian Type I error." See Morton (1955) for a use of this phrase, as well as a similar development of this concept in the context of genetic linkage analysis. Whereas the FWER is very much frequentist, we have shown that the pFDR is quite flexible in its interpretation. This is especially appealing in that it is a multiple testing error measure that can be used by both Bayesians and frequentists. We will see in later examples that this is easily accomplished.

The quantity $\mathrm{pFDR}(\Gamma) = \Pr(H = 0 | T \in \Gamma)$ gives a global measure in that it does not provide specific information about the value of each statistic, only whether it fell in Γ or not. In Section 4, we use the pFDR to give each statistic a measure of its significance in terms of the pFDR, which we call the q-value. This continues to have a Bayesian interpretation, yet allows one to make simultaneous inferences. Reporting marginal posterior probabilities $\Pr(H = 0 | T = t)$, as is the case in typical Bayesian modeling, also gives a specific measure for each statistic, but it does not take into account the multiple comparisons.

4. The q-value. In this section, we introduce the pFDR analogue of the p-value, which we call the *q-value*. Because of the connection made in Section 3, the q-value is useful in both Bayesian and frequentist settings. It gives the scientist a hypothesis testing error measure for each observed statistic with respect to the pFDR. Again, assume that (T_i, H_i) are i.i.d. random variables, $T_i|H_i \sim (1 - H_i) \cdot F_0 + H_i \cdot F_1$, and $H_i \sim \text{Bernoulli}(1 - \pi_0)$ for $i = 1, \ldots, m$. We introduce the q-value by first showing an example.

EXAMPLE [Testing the mean of a $N(\theta, 1)$ random variable]. Suppose we perform m hypothesis tests of $\theta = 0$ versus $\theta = 2$ for m $N(\theta, 1)$ random variables $T_1, \ldots, T_m$. Specifically, (T_i, H_i) are i.i.d. random variables with $T_i|H_i \sim (1 - H_i) \cdot N(0, 1) + H_i \cdot N(2, 1)$. Given we observe the random variables to be $T_1 = t_1, \ldots, T_m = t_m$, the p-value of $T_i = t_i$ can be calculated as

$$p\text{-value}(t_i) = \Pr(T \geq t_i | H = 0) = \Pr\big(N(0, 1) \geq t_i\big).$$

In words, p-value(t_i) gives the Type I error rate if we reject any statistic as extreme or more extreme than t_i.

By Theorem 1 the pFDR, if we reject any statistic as extreme or more extreme than t_i, among all m hypotheses is

$$\begin{aligned} \text{pFDR}(\{T \geq t_i\}) &= \frac{\pi_0 \Pr(T \geq t_i | H = 0)}{\pi_0 \Pr(T \geq t_i | H = 0) + \pi_1 \Pr(T \geq t_i | H = 1)} \\ &= \frac{\pi_0 \Pr(N(0, 1) \geq t_i)}{\pi_0 \Pr(N(0, 1) \geq t_i) + \pi_1 \Pr(N(2, 1) \geq t_i)} \\ &= \Pr(H = 0 | T \geq t_i). \end{aligned} \tag{4.1}$$

From the last line, it can be seen that $\text{pFDR}(\{T \geq t_i\}) = \Pr(H = 0|T \geq t_i)$ is a natural Bayesian analogue to p-value$(t_i) = \Pr(T \geq t_i|H = 0)$. The relationship between these two quantities can also be understood graphically. Figure 1 shows a graph of the $N(0, 1)$ and $N(2, 1)$ distributions with the point $T_i = t_i$ marked by a vertical line. The area under the $N(0, 1)$ density to the right of t_i is p-value(t_i). In order to calculate $\text{pFDR}(\{T \geq t_i\})$, we use formula (4.1), which involves the areas to the right of t_i under the $N(0, 1)$ and the $N(2, 1)$ densities, and their respective weights π_0 and π_1.

As we show in this section, (4.1) is what we call q-value(t_i). In many situations, it is the pFDR obtained when rejecting a statistic as extreme or more extreme than t_i among all m hypotheses; but the q-value can be defined more generally, as can the p-value.

Until now, we have only considered a single significance region. Hypothesis tests are usually derived according to a nested set of significance regions. As long as F_0 and F_1 have a common support, we can denote this nested set of significance regions without loss of generality by $\{\Gamma_\alpha\}_{\alpha=0}^{1}$, where α is such that

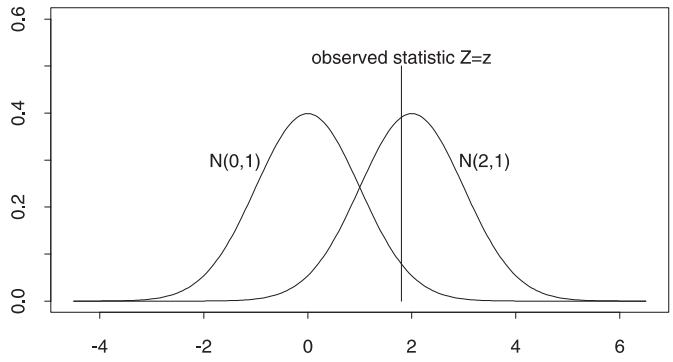

FIG. 1. *A plot of the* $N(0,1)$ *and* $N(2,1)$ *densities. The vertical line denotes the observed statistic* $T_i = t_i$. *The* p*-value can be calculated from the area under the* $N(0,1)$ *density to the right of* $T_i = t_i$. *The* q*-value is calculated using the area under both densities to the right of* $T_i = t_i$ *weighted by* π_0 *and* π_1.

$\Pr(T \in \Gamma_\alpha | H = 0) = \alpha$. Note that $\alpha' \leq \alpha$ implies that $\Gamma_{\alpha'} \subseteq \Gamma_\alpha$, giving the nested property. Using this notation, the p-value(t) of an observed statistic $T = t$ is defined to be [Lehmann (1986)]

$$p\text{-value}(t) = \inf_{\{\Gamma_\alpha : t \in \Gamma_\alpha\}} \Pr(T \in \Gamma_\alpha | H = 0).$$

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 t, given the set of nested significance regions.

We define an analogous quantity in terms of the pFDR that has both frequentist multiple hypothesis testing and Bayesian interpretations.

DEFINITION 2. For an observed statistic $T = t$ define the *q-value* of t to be

$$q\text{-value}(t) = \inf_{\{\Gamma_\alpha : t \in \Gamma_\alpha\}} \text{pFDR}(\Gamma_\alpha). \tag{4.2}$$

In words, (4.2) says the q-value is a measure of the strength of an observed statistic with respect to the pFDR; it is the minimum pFDR that can occur when rejecting a statistic with value t for the set of nested significance regions. Under the assumptions of Theorem 1, it is seen that the q-value has an even more interpretable relationship to the p-value. Consider the following corollary.

COROLLARY 2. *Under the assumptions of Theorem* 1,

$$q\text{-value}(t) = \inf_{\{\Gamma_\alpha : t \in \Gamma_\alpha\}} \Pr(H = 0 | T \in \Gamma_\alpha). \tag{4.3}$$

Therefore, according to (4.3), the q-value is a Bayesian version of the p-value—say a "posterior Bayesian p-value"—the minimum posterior probability $H = 0$

over all significance regions containing the statistic. We call (4.3) the q-value because it is equivalent to the p-value with the events $\{T \in \Gamma_\alpha\}$ and $\{H = 0\}$ reversed.

REMARK. Technically, the q-value is not a "pFDR adjusted p-value." This is immediately clear when recalling the definition of an adjusted p-value. Shaffer (1995) says, "Given any test procedure, the adjusted p-value corresponding to a test of a single hypothesis H_i can be defined as the level of the entire test procedure at which H_i would be rejected, given the values of all test statistics involved." Therefore, since the pFDR cannot be controlled by a test procedure (i.e., a sequential p-value method), then it cannot be used to define any adjusted p-values. But more importantly, notice that the adjusted p-values are defined *in terms of a particular procedure*.

Notice that

$$\begin{aligned}
&\underset{\{\Gamma_\alpha : t \in \Gamma_\alpha\}}{\arg\min}\ \Pr(H = 0 | T \in \Gamma_\alpha) \\
&\quad = \underset{\{\Gamma_\alpha : t \in \Gamma_\alpha\}}{\arg\min}\ \frac{\pi_0 \Pr(T \in \Gamma_\alpha | H = 0)}{\pi_0 \Pr(T \in \Gamma_\alpha | H = 0) + \pi_1 \Pr(T \in \Gamma_\alpha | H = 1)} \\
&\quad = \underset{\{\Gamma_\alpha : t \in \Gamma_\alpha\}}{\arg\min}\ \frac{\Pr(T \in \Gamma_\alpha | H = 0)}{\Pr(T \in \Gamma_\alpha | H = 1)}.
\end{aligned}$$

Therefore, the significance region that determines the q-value minimizes the ratio of the Type I error to the power over all significance regions that contain the statistic. This makes sense because the pFDR is concerned with measuring how frequently the false positives occur in relation to true positives.

One can understand this observation in terms of a plot of power versus Type I error for a given set of significance regions. Recall that marginally $T_i \overset{\text{i.i.d.}}{\sim} \pi_0 \cdot F_0 + \pi_1 \cdot F_1$ for $i = 1, \ldots, m$. We write

$$\begin{aligned}
G_1(\alpha) &:= \int_{\Gamma_\alpha} dF_1 = \Pr(T \in \Gamma_\alpha | H = 1), \\
G_0(\alpha) &:= \int_{\Gamma_\alpha} dF_0 = \Pr(T \in \Gamma_\alpha | H = 0) = \alpha.
\end{aligned} \tag{4.4}$$

It is easily shown that G_0 and G_1 are the c.d.f.'s of the null and alternative p-values, respectively. Suppose that G_1 is continuous and differentiable. Then it can be shown through simple calculus that $\alpha / G_1(\alpha)$ is minimized at $\alpha = G_1(\alpha) / G_1'(\alpha)$. Therefore we can minimize $\alpha / G_1(\alpha)$ graphically by drawing all lines from the origin that are tangent to a concave portion of the function. The line with the largest slope is tangent to the point on the curve where $\alpha / G_1(\alpha)$ is minimized.

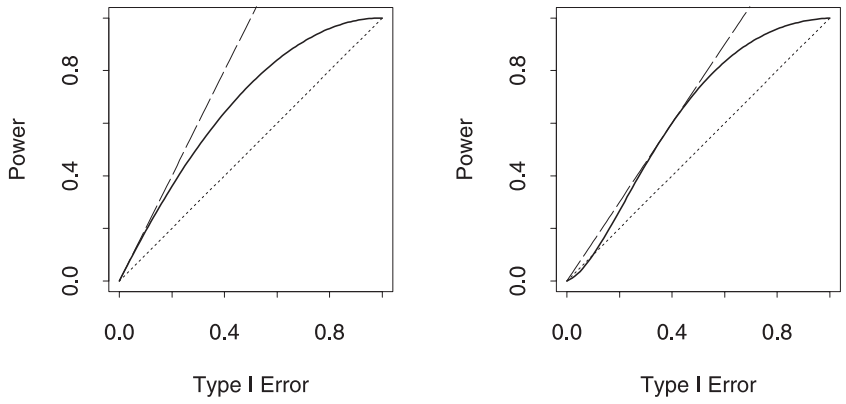

FIG. 2. *A plot of power versus Type* I *error rate for two hypothetical sets of significance regions. The solid line is power as a function of Type* I *error,* $G_1(\alpha)$; *the dotted line is the identity function*; *the dashed line is the line from the origin tangent to* $G_1(\alpha)$.

See Figure 2 for a picture of this maximization. The left panel has a strictly concave $G_1(\alpha)$. In this case, the ratio of power to Type I errors decreases as $\alpha \to 0$. In other words, as the significance regions get smaller, the ratio of power to Type I errors gets larger. Therefore, for a concave G_1, we can conclude that the Γ_α that contains t and minimizes $\text{pFDR}(\Gamma_\alpha)$ also minimizes $\Pr(T \in \Gamma_\alpha | H = 0)$. This follows since we would take the significance region with the smallest α, where $t \in \Gamma_\alpha$, in order to minimize $\alpha / G_1(\alpha)$.

Therefore, when the c.d.f. of the alternative p-values G_1 is concave, the same significance region is used to define the q-value and the p-value. More generally, we only need that $G_1(\alpha)/\alpha$ is a decreasing function of α if we do not assume that G_1 is differentiable. (Note that if G_1 is concave then this holds.) We state this formally in the following proposition.

PROPOSITION 1. *The* q*-value of a statistic is based on the same significance region as the* p*-value, as long as* $G_1(\alpha)/\alpha$ *is decreasing in* α, *that is*,

$$\underset{\{\Gamma_\alpha : t \in \Gamma_\alpha\}}{\arg\min}\ \Pr(H = 0 | T \in \Gamma_\alpha) = \underset{\{\Gamma_\alpha : t \in \Gamma_\alpha\}}{\arg\min}\ \Pr(T \in \Gamma_\alpha | H = 0).$$

The right panel of Figure 2 shows an example where G_1 is not concave, nor is $G_1(\alpha)/\alpha$ decreasing in α. The significance region minimizing the ratio of the Type I error to the power is the one that corresponds to the point that the shown line from the origin intersects. No similar connection can be made with the p-value under this kind of G_1.

EXAMPLE (Likelihood ratio based rejection regions). We have assumed that (T_i, H_i) are i.i.d. random variables where $T_i | H_i \sim (1 - H_i) \cdot F_0 + H_i \cdot F_1$. Suppose

that f_0 and f_1 are the densities corresponding to F_0 and F_1, respectively. Also, suppose that we consider significance regions of the form

$$\left\{t : \frac{f_0(t)}{f_1(t)} \leq \lambda\right\}.$$

Then it follows that the power to Type I error curve is concave, and therefore Proposition 1 holds. Moreover, the result of Theorem 2 that we present below also holds.

It is natural to consider whether defining the q-value in terms of the original statistics is equivalent to defining the q-value in terms of the statistics' p-values. We denote the pFDR based on the original statistics as $\text{pFDR}^T(\Gamma_\alpha)$, and the analogous pFDR based on the p-values by $\text{pFDR}^P(\{p \leq \alpha\})$. Is $\text{pFDR}^T(\Gamma_\alpha) = \text{pFDR}^P(\{p \leq \alpha\})$, and when is it the case that $q\text{-value}(t_i) = \text{pFDR}^P(\{p \leq p\text{-value}(t_i)\})$? We answer these questions in the following theorem.

THEOREM 2. *For m identical hypothesis tests, $\text{pFDR}^T(\Gamma_\alpha) = \text{pFDR}^P(\{p : p \leq \alpha\})$, which implies that the q-value can be calculated from either the original statistics or their p-values. Also, when the statistics are independent and follow a mixture distribution then*

$$q\text{-value}(t) = \text{pFDR}^P\big(\{p : p \leq p\text{-value}(t)\}\big)$$

if and only if $G_1(\alpha)/\alpha$ is decreasing in α.

PROOF. Because the set of significance regions is nested, it is trivial to show that $p\text{-value}(t) \leq \alpha$ if and only if $t \in \Gamma_\alpha$. This implies $\text{pFDR}^T(\Gamma_\alpha) = \text{pFDR}^P(\{p : p \leq \alpha\})$. For the second statement, first suppose that $G_1(\alpha)/\alpha$ is decreasing in α. For any $T = t$, let $\Gamma_{\alpha'} = \arg\min_{\{\Gamma_\alpha : t \in \Gamma_\alpha\}} \text{pFDR}^T(\Gamma_\alpha)$, so that $q\text{-value}(t) = \text{pFDR}^T(\Gamma_{\alpha'}) = \text{pFDR}^P(\{p : p \leq \alpha'\})$. It is also the case that $\Gamma_{\alpha'} = \arg\min_{\{\Gamma_\alpha : t \in \Gamma_\alpha\}} \Pr(T \in \Gamma_\alpha | H = 0)$, that is, $p\text{-value}(t) = \alpha'$. Now suppose that $q\text{-value}(t) = \text{pFDR}(\{p : p \leq p\text{-value}(t)\})$ for each t. By the definition of the q-value, this implies that $q\text{-value}(t)$ is an increasing function of $p\text{-value}(t)$. Therefore $G_1(\alpha)/\alpha$ is a decreasing function of α. □

Suppose that we perform m different hypothesis tests, so that each one has its own nested set of significance regions, possibly on different spaces. One can transform these tests into the same space by calculating their p-values. By the results presented in the latter half of this section, it follows that the p-value based q-values are a natural way to transform these tests onto the same space with respect to the pFDR. Methods have also been developed for estimating q-values [Storey (2002a)], and Storey, Taylor and Siegmund (2004) show that these estimates are simultaneously conservatively consistent.

5. Dependence and asymptotic properties. In this section, we consider the pFDR when the test statistics are dependent, along with some asymptotic properties that have direct applications to certain cases of dependence. Here we assume we are performing m hypothesis tests based on statistics $T_1, \ldots, T_m$ and using the same significance region for each. We first present the following simple result.

THEOREM 3. *Under any distributional assumptions about $T_1, \ldots, T_m$ and $H_1, \ldots, H_m$, it follows that*

$$\mathrm{pFDR}(\Gamma) = \sum_{k=1}^{m} \sum_{i_1,\ldots,i_k} \frac{1}{k} \sum_{j=1}^{k} \Pr\left(H_{i_j} = 0, \begin{array}{c} T_{i_1}, \ldots, T_{i_k} \in \Gamma \\ T_{i_{k+1}}, \ldots, T_{i_m} \notin \Gamma \end{array} \middle| R(\Gamma) > 0 \right),$$

where the middle sum is taken over all distinct subsets of size k of $\{1, 2, \ldots, m\}$.

PROOF.

$$\begin{aligned}
\mathrm{pFDR}(\Gamma) &= \sum_{k=1}^{m} \sum_{i_1,\ldots,i_k} \mathrm{E}\left(\frac{V(\Gamma)}{R(\Gamma)} \middle| \begin{array}{c} T_{i_1}, \ldots, T_{i_k} \in \Gamma \\ T_{i_{k+1}}, \ldots, T_{i_m} \notin \Gamma \end{array} \right) \\
&\qquad \times \Pr\left(\begin{array}{c} T_{i_1}, \ldots, T_{i_k} \in \Gamma \\ T_{i_{k+1}}, \ldots, T_{i_m} \notin \Gamma \end{array} \middle| R(\Gamma) > 0 \right) \\
&= \sum_{k=1}^{m} \sum_{i_1,\ldots,i_k} \mathrm{E}\left(\frac{\sum_{j=1}^{k} (1 - H_{i_j})}{k} \middle| \begin{array}{c} T_{i_1}, \ldots, T_{i_k} \in \Gamma \\ T_{i_{k+1}}, \ldots, T_{i_m} \notin \Gamma \end{array} \right) \\
&\qquad \times \Pr\left(\begin{array}{c} T_{i_1}, \ldots, T_{i_k} \in \Gamma \\ T_{i_{k+1}}, \ldots, T_{i_m} \notin \Gamma \end{array} \middle| R(\Gamma) > 0 \right) \\
&= \sum_{k=1}^{m} \sum_{i_1,\ldots,i_k} \frac{1}{k} \sum_{j=1}^{k} \Pr\left(H_{i_j} = 0 \middle| \begin{array}{c} T_{i_1}, \ldots, T_{i_k} \in \Gamma \\ T_{i_{k+1}}, \ldots, T_{i_m} \notin \Gamma \end{array} \right) \\
&\qquad \times \Pr\left(\begin{array}{c} T_{i_1}, \ldots, T_{i_k} \in \Gamma \\ T_{i_{k+1}}, \ldots, T_{i_m} \notin \Gamma \end{array} \middle| R(\Gamma) > 0 \right). \qquad \square
\end{aligned}$$

The representation of $\mathrm{pFDR}(\Gamma)$ in Theorem 3 appears intractable at first glance, but under a fully parametric model it is feasible to calculate this quantity or a numerical approximation to it. When the tests are exchangeable but dependent in some arbitrary way, we may simplify this result.

COROLLARY 3. *Suppose that $(H_1, T_1), \ldots, (H_m, T_m)$ are exchangeable random variables. Then*

$$\mathrm{pFDR}(\Gamma_\alpha) = \sum_{k=1}^{m} \Pr\left(H_1 = 0 \middle| \begin{array}{c} T_1, \ldots, T_k \in \Gamma_\alpha \\ T_{k+1}, \ldots, T_m \notin \Gamma_\alpha \end{array} \right) \cdot \Pr(R = k | R > 0).$$

From these results it can be seen that Theorem 1 does not hold under general dependence. Therefore, we now determine when Theorem 1 holds approximately, or rather asymptotically. Recall that we can represent the nested set of significance regions by $\{\Gamma_\alpha\}_{\alpha>0}$, where α is the Type I error of Γ_α. In our notation,

$$\frac{V_m(\Gamma_\alpha)}{\sum_{i=1}^m (1-H_i)} = \frac{\sum_{i=1}^m (1-H_i)\cdot \mathbb{1}(T_i \in \Gamma_\alpha)}{\sum_{i=1}^m (1-H_i)},$$
$$\frac{S_m(\Gamma_\alpha)}{\sum_{i=1}^m H_i} = \frac{\sum_{i=1}^m H_i \cdot \mathbb{1}(T_i \in \Gamma_\alpha)}{\sum_{i=1}^m H_i}$$

represent the empirical distribution functions of the null and alternative p-values, respectively, as a function of α. If these quantities converge in the pointwise sense, then $\mathrm{pFDR}(\Gamma_\alpha)$, $\mathrm{FDR}(\Gamma_\alpha)$ and $V(\Gamma_\alpha)/R(\Gamma_\alpha)$ all converge to a posterior probability, simultaneously for all Γ_α. This is explicitly stated in the following theorem.

THEOREM 4. *Suppose that with probability* 1 *we have*

$$\sum_{i=1}^m (1-H_i)/m \to \pi_0$$

and

$$\frac{V_m(\Gamma_\alpha)}{\sum_{i=1}^m (1-H_i)} \to G_0(\alpha), \qquad \frac{S_m(\Gamma_\alpha)}{\sum_{i=1}^m H_i} \to G_1(\alpha),$$

for each $\alpha > 0$ *for some continuous functions* G_0 *and* G_1, *as* $m \to \infty$. *Then for any* $\delta > 0$ *where* $\pi_0 \cdot G_0(\delta) + (1-\pi_0)\cdot G_1(\delta) > 0$,

(i) $$\lim_{m\to\infty} \sup_{\alpha\geq\delta} \left| \frac{V_m(\Gamma_\alpha)}{R_m(\Gamma_\alpha)\vee 1} - \Pr_\infty(H=0|X\in\Gamma_\alpha) \right| \overset{a.s.}{=} 0,$$

(ii) $$\lim_{m\to\infty} \sup_{\alpha\geq\delta} \left| \mathrm{FDR}_m(\Gamma_\alpha) - \Pr_\infty(H=0|X\in\Gamma_\alpha) \right| = 0,$$

(iii) $$\lim_{m\to\infty} \sup_{\alpha\geq\delta} \left| \mathrm{pFDR}_m(\Gamma_\alpha) - \Pr_\infty(H=0|X\in\Gamma_\alpha) \right| = 0,$$

where we define

$$\Pr_\infty(H=0|X\in\Gamma_\alpha) = \frac{\pi_0\cdot G_0(\alpha)}{\pi_0\cdot G_0(\alpha) + (1-\pi_0)\cdot G_1(\alpha)}.$$

The functions G_0 and G_1 are the asymptotic Type I error and power of the p-values as a function of α. In general, Theorem 4 says that if the statistics are "weakly dependent" then the realized proportion of false discoveries, the FDR, and the pFDR converge simultaneously over all significance regions to the Bayesian posterior probability defined above. It allows many of the properties shown for the q-value under independence to hold approximately under weak dependence.

Most importantly, Theorem 4 says that if one is able to calculate or estimate G_0, G_1 and π_0, then for large m these provide good approximations for the realized proportion of false discoveries, the FDR, and the pFDR for all significance regions simultaneously.

Two useful cases where the theorem may hold are when the statistics $T_1, T_2, \ldots$ are such that there exists a k where $|i-j| \geq k$ implies T_i and T_j are independent (i.e., the dependence is in finite blocks), or when the statistics are a stationary ergodic sequence. There are other forms of dependence where this result holds, for example, for certain Markov chains and certain mixing distributions.

PROOF OF THEOREM 4. Let

$$Q_m(\alpha) = \frac{V_m(\Gamma_\alpha)}{[V_m(\Gamma_\alpha) + S_m(\Gamma_\alpha)] \vee 1}.$$

By an easy modification of the Glivenko–Cantelli theorem [see, e.g., Billingsley (1968)], it follows that

$$\lim_{m\to\infty} \sup_{\alpha\geq\delta} \left| \frac{V_m(\Gamma_\alpha)}{m} - \pi_0 \cdot G_0(\alpha) \right| \overset{\text{a.s.}}{=} 0,$$

$$\lim_{m\to\infty} \sup_{\alpha\geq\delta} \left| \frac{[V_m(\Gamma_\alpha) + S_m(\Gamma_\alpha)] \vee 1}{m} - [\pi_0 \cdot G_0(\alpha) + \pi_1 \cdot G_1(\alpha)] \right| \overset{\text{a.s.}}{=} 0.$$

Since $\pi_0 \cdot G_0(\delta) + (1-\pi_0) \cdot G_1(\delta) > 0$ and these are both nondecreasing functions, it is easy to show that

$$\lim_{m\to\infty} \sup_{\alpha\geq\delta} \left| \frac{m}{[V_m(\Gamma_\alpha) + S_m(\Gamma_\alpha)] \vee 1} - \frac{1}{\pi_0 \cdot G_0(\alpha) + \pi_1 \cdot G_1(\alpha)} \right| \overset{\text{a.s.}}{=} 0.$$

Finally, noticing that

$$\begin{aligned} &\left| \frac{V_m(\Gamma_\alpha) - m\pi_0 \cdot G_0(\alpha)}{[V_m(\Gamma_\alpha) + S_m(\Gamma_\alpha)] \vee 1} \right| \\ &\quad + \left| \frac{m\pi_0 \cdot G_0(\alpha)}{[V_m(\Gamma_\alpha) + S_m(\Gamma_\alpha)] \vee 1} - \frac{\pi_0 \cdot G_0(\alpha)}{\pi_0 \cdot G_0(\alpha) + \pi_1 \cdot G_1(\alpha)} \right| \\ &\quad \geq \left| Q_m(\alpha) - \Pr_\infty(H = 0 | X \in \Gamma_\alpha) \right| \end{aligned}$$

completes the proof of the first convergence.

Now $|Q_m(\alpha) - \Pr_\infty(H = 0|X \in \Gamma_\alpha)| \leq 2$ almost surely, so that it easily follows that

$$\begin{aligned} 0 &= \mathrm{E}\left[\lim_{m\to\infty} \sup_{\alpha\geq\delta} \left| Q_m(\alpha) - \Pr_\infty(H = 0|X \in \Gamma_\alpha) \right| \right] \\ &= \lim_{m\to\infty} \mathrm{E}\left[\sup_{\alpha\geq\delta} \left| Q_m(\alpha) - \Pr_\infty(H = 0|X \in \Gamma_\alpha) \right| \right] \\ &\geq \lim_{m\to\infty} \sup_{\alpha\geq\delta} \left| \mathrm{E}[Q_m(\alpha)] - \Pr_\infty(H = 0|X \in \Gamma_\alpha) \right| \geq 0, \end{aligned}$$

where $\mathrm{E}[Q_m(\alpha)] = \mathrm{FDR}_m(\Gamma_\alpha)$. Finally,

$$\lim_{m\to\infty} \sup_{\alpha\geq\delta} |\,\mathrm{pFDR}_m(\Gamma_\alpha) - \mathrm{FDR}_m(\Gamma_\alpha)| \leq \lim_{m\to\infty} \left| \frac{1}{\Pr(R_m(\delta) > 0)} - 1 \right| = 0. \qquad \square$$

The following is a numerical example involving statistics that are dependent in finite blocks. A specific application where this type of dependence plausibly exists is discussed in Section 7. There, hypotheses are tested on thousands of genes' expression levels from several biological samples. Since genes tend to work in pathways of finite size, it is likely that the dependence between the genes exists in relatively small, disjoint groups. Thus, convergence of the quantities in Theorem 4 likely occurs in that application. For more on this, see Storey (2002b).

EXAMPLE (Locally dependent statistics). As a numerical example to illustrate the result of Theorem 4, consider the following situation. Suppose $T_i|H_i = 0 \sim N(0,1)$ and $T_i|H_i = 1 \sim N(2,1)$. We have $\mathrm{Cov}(T_{i+j}, T_{i+k}) = \rho$ where $0 \leq \rho \leq 1$ for $j \neq k$, $j, k = 0, 1, \ldots, 9$ and $i = 1, 11, 21, \ldots$, and zero covariance otherwise. In other words the statistics have correlation ρ in groups of 10. Suppose we take $\Gamma_\alpha = [\Phi^{-1}(1-\alpha), \infty)$ where Φ is the c.d.f. of a $N(0,1)$, and $H_i \overset{\text{i.i.d.}}{\sim}$ Bernoulli$(1-\pi_0)$, with $\pi_0 = 0.8$. Then by Theorem 4, we have, for example, that

$$\lim_{m\to\infty} \sup_{\alpha>0} |\mathrm{pFDR}_m(\Gamma_\alpha) - \Pr_\infty(H = 0|T \in \Gamma_\alpha)| = 0,$$

where $\Pr_\infty(H = 0|T \in \Gamma_\alpha) = \pi_0 \cdot \alpha / [\pi_0 \cdot \alpha + (1-\pi_0) \cdot \Pr(N(2,1) \geq \Phi^{-1}(1-\alpha))]$. Table 2 shows $\Pr_\infty(H = 0|T \in \Gamma_\alpha)$ compared to the $\mathrm{pFDR}_m(\Gamma_\alpha)$ at

TABLE 2
Simulation results: $\mathrm{pFDR}_m(\Gamma_\alpha)$ *converging to* $\Pr_\infty(H = 0|T \in \Gamma_\alpha)$

m	$\rho = 0$	$\rho = 0.2$	$\rho = 0.4$	$\rho = 0.6$	$\rho = 0.8$	$\rho = 1$
$\alpha = 0.005$, $\Pr_\infty(H = 0\|T \in \Gamma_\alpha) = 0.137$						
100	0.142 (0.004)	0.126 (0.004)	0.120 (0.004)	0.102 (0.004)	0.094 (0.004)	0.041 (0.003)
500	0.136 (0.003)	0.136 (0.003)	0.133 (0.003)	0.127 (0.003)	0.117 (0.003)	0.091 (0.003)
1000	0.138 (0.003)	0.136 (0.003)	0.134 (0.003)	0.132 (0.003)	0.128 (0.003)	0.113 (0.003)
3000	0.138 (0.003)	0.137 (0.003)	0.137 (0.003)	0.137 (0.003)	0.134 (0.003)	0.129 (0.003)
5000	0.138 (0.003)	0.138 (0.003)	0.137 (0.003)	0.137 (0.003)	0.135 (0.003)	0.132 (0.003)
$\alpha = 0.001$, $\Pr_\infty(H = 0\|T \in \Gamma_\alpha) = 0.061$						
100	0.061 (0.003)	0.063 (0.004)	0.053 (0.003)	0.047 (0.003)	0.036 (0.003)	0.010 (0.002)
500	0.061 (0.002)	0.063 (0.002)	0.060 (0.002)	0.052 (0.002)	0.047 (0.002)	0.028 (0.002)
1000	0.063 (0.002)	0.062 (0.002)	0.060 (0.002)	0.060 (0.002)	0.055 (0.002)	0.038 (0.002)
3000	0.061 (0.001)	0.063 (0.001)	0.061 (0.001)	0.061 (0.001)	0.058 (0.001)	0.051 (0.002)
5000	0.061 (0.001)	0.063 (0.001)	0.062 (0.001)	0.062 (0.001)	0.060 (0.001)	0.054 (0.002)

several m for $\alpha = 0.005$ and $\alpha = 0.001$. It can be seen that there is quite good agreement between the limiting case and the finite cases, especially for large m. Most of the differences at $m = 5000$ are within the Monte Carlo standard error, which is listed parenthetically.

6. A connection to classification theory. When assuming the statistics follow a mixture distribution, as we have assumed throughout this work, it is possible to view multiple hypothesis testing as a classification problem. For each test, we observe T_i and we have to decide whether to classify H_i as 0 or H_i as 1 based on T_i. There are four possible outcomes for each test with two of them being misclassifications. Consider Table 3 listing these outcomes, with the penalties for each type of misclassification parameterized by λ.

We use several of the basic facts about classification theory found in Cherkassky and Mulier (1998), for example. A significance region Γ can be thought of as a classification rule in the following way: if $T_i \in \Gamma$ then we classify H_i as 1, and if $T_i \notin \Gamma$, then we classify H_i as 0. The "Bayes error" of a classification rule (in terms of the significance region representation) is

$$\mathrm{BE}(\Gamma) = (1-\lambda)\Pr(T_i \in \Gamma, H_i = 0) + \lambda \Pr(T_i \notin \Gamma, H_i = 1). \tag{6.1}$$

That is, $\mathrm{BE}(\Gamma)$ is the expected loss under Table 3.

Genovese and Wasserman (2002a) notice that one can define a dual quantity to the FDR, which they call the false nondiscovery rate (FNR). [See also Sarkar (2002).] The FNR is defined to be the expected proportion of false negatives among all hypotheses that *are not* rejected, with the ratio being set to zero if all hypotheses are rejected:

$$\mathrm{FNR} = \mathrm{E}\left[\frac{T}{W}\middle| W > 0\right]\Pr(W > 0), \tag{6.2}$$

where W is the total number of nonsignificant hypotheses, and T (not to be confused with the statistics T_i) is the number of nonsignificant alternative statistics (false negatives). We make the following modified definition of the FNR, in the spirit of the pFDR.

DEFINITION 3. The *positive false nondiscovery rate* is defined to be:

$$\mathrm{pFNR} = \mathrm{E}\left[\frac{T}{W}\middle| W > 0\right].$$

Using an analogous argument to Theorem 1, we can show the following result.

TABLE 3
Outcomes of "classifying" H_i with misclassification penalties

	Classify H_i as 0	**Classify H_i as 1**
$H_i = 0$	0	$1-\lambda$
$H_i = 1$	λ	0

THEOREM 5. *Under the assumptions of Theorem* 1, *it follows that*

$$\text{pFNR}(\Gamma) = \Pr(H = 1 | T \notin \Gamma),$$

where $\pi_1 = 1 - \pi_0$ *is the implicit prior probability in the above posterior probability.*

Now the Bayes error can be written as a weighted sum of pFDR(Γ) and pFNR(Γ).

COROLLARY 4. *Under the assumptions of Theorem* 1,

$$\text{BE}(\Gamma) = (1 - \lambda) \Pr(T \in \Gamma) \cdot \text{pFDR}(\Gamma) + \lambda \Pr(T \notin \Gamma) \cdot \text{pFNR}(\Gamma). \tag{6.3}$$

In the step-wise p-value framework, one decides beforehand at what level to control the FDR and then applies one's procedure to control it at that level. Using the classification theory connection, we suggest two ways to use the pFDR in fixing a significance region beforehand. One can choose the significance region based on the relative cost of a false positive to a false negative and then minimize the Bayes error; or one can decide the relative importance of the pFDR to the pFNR and then minimize their weighted average.

In Section 7, we consider a problem in which one is concerned with deciding which of several thousand genes show a statistically significant change in gene expression between two types of cells (e.g., normal versus diseased cells). Here it is feasible that the scientist can decide on the relative cost of a false positive gene to a false negative gene. In that case, one can derive the Bayes rule to minimize the Bayes error. By Corollary 4, one can interpret the Bayes error in terms of the multiple hypothesis testing quantities pFDR and pFNR. In fact, the manner in which the Bayes error weights the pFDR and pFNR makes a lot of sense. Another approach is to minimize the weighted average of the pFDR and pFNR,

$$(1 - w) \cdot \text{pFDR}(\Gamma) + w \cdot \text{pFNR}(\Gamma).$$

In words, one can decide how important the rate of false discoveries is to the rate of false nondiscoveries. We now show how to minimize this weighted average.

Recall that we assume (T_i, H_i) are i.i.d. random variables, $T_i | H_i \sim (1 - H_i) \cdot F_0 + H_i \cdot F_1$, and $H_i \sim \text{Bernoulli}(1 - \pi_0)$ for $i = 1, \ldots, m$. Also assume that F_0 and F_1 are continuous distributions with common support, with respective densities f_0 and f_1. Define the set of significance regions $\{\mathcal{B}_\lambda\}$ for $0 \leq \lambda \leq 1$ by

$$\mathcal{B}_\lambda = \left\{ t : \frac{\pi_0 f_0(t)}{\pi_0 f_0(t) + \pi_1 f_1(t)} \leq \lambda \right\}.$$

The set $\{\mathcal{B}_\lambda\}$ defines the Bayes rule for the cost matrix given by Table 3. That is, for each λ, $\mathcal{B}_\lambda$ minimizes BE($\mathcal{B}_\lambda$) (6.1). Note that by Corollary 4, $\mathcal{B}_\lambda$ also minimizes (6.3) for each λ.

As it turns out, the nested set of significance regions $\{\mathcal{B}_\lambda\}$ can also be used to minimize $(1-w)\cdot \mathrm{pFDR}(\Gamma) + w \cdot \mathrm{pFNR}(\Gamma)$. We state this formally in the following theorem.

THEOREM 6. *Let* $\lambda(w) = \arg\min_\lambda[(1-w)\cdot \mathrm{pFDR}(\mathcal{B}_\lambda) + w \cdot \mathrm{pFNR}(\mathcal{B}_\lambda)]$. *Then* $(1-w)\cdot \mathrm{pFDR}(\mathcal{B}_{\lambda(w)}) + w \cdot \mathrm{pFNR}(\mathcal{B}_{\lambda(w)})$ *minimizes* $(1-w)\cdot \mathrm{pFDR}(\Gamma) + w \cdot \mathrm{pFNR}(\Gamma)$ *among all measurable* Γ.

PROOF. Recall that by the Neyman–Pearson lemma, the $\{\mathcal{B}_\lambda\}$ form a set of uniformly most powerful significance regions. Without loss of generality, we can assume that, for each $\alpha \in [0,1]$, there exists a $\mathcal{B}_\lambda$ such that $\Pr(T \in \mathcal{B}_\lambda | H = 0) = \alpha$. Otherwise, $\{\mathcal{B}_\lambda\}$ can be extended in the natural way to accomplish this and still remain uniformly most powerful [Lehmann (1986)].

Consider any measurable Γ. Then there exists a $\mathcal{B}_\lambda$ such that $\Pr(T \in \Gamma | H = 0) = \Pr(T \in \mathcal{B}_\lambda | H = 0)$. Since the $\{\mathcal{B}_\lambda\}$ are uniformly most powerful, it follows that $\Pr(T \in \Gamma | H = 1) \le \Pr(T \in \mathcal{B}_\lambda | H = 1)$. Therefore,

$$\begin{aligned}
\mathrm{pFDR}(\Gamma) &= \frac{\pi_0 \cdot \Pr(T \in \Gamma | H = 0)}{\pi_0 \cdot \Pr(T \in \Gamma | H = 0) + \pi_1 \cdot \Pr(T \in \Gamma | H = 1)} \\
&\ge \frac{\pi_0 \cdot \Pr(T \in \mathcal{B}_\lambda | H = 0)}{\pi_0 \cdot \Pr(T \in \mathcal{B}_\lambda | H = 0) + \pi_1 \cdot \Pr(T \in \mathcal{B}_\lambda | H = 1)} = \mathrm{pFDR}(\mathcal{B}_\lambda), \\
\mathrm{pFNR}(\Gamma) &= \frac{\pi_1 \cdot \Pr(T \notin \Gamma | H = 1)}{\pi_1 \cdot \Pr(T \notin \Gamma | H = 1) + \pi_0 \cdot \Pr(T \notin \Gamma | H = 0)} \\
&\ge \frac{\pi_1 \cdot \Pr(T \notin \mathcal{B}_\lambda | H = 1)}{\pi_1 \cdot \Pr(T \notin \mathcal{B}_\lambda | H = 1) + \pi_0 \cdot \Pr(T \notin \mathcal{B}_\lambda | H = 0)} = \mathrm{pFNR}(\mathcal{B}_\lambda).
\end{aligned}$$

Hence for any w, $(1-w)\cdot \mathrm{pFDR}(\mathcal{B}_\lambda) + w \cdot \mathrm{pFNR}(\mathcal{B}_\lambda) \le (1-w)\cdot \mathrm{pFDR}(\Gamma) + w \cdot \mathrm{pFNR}(\Gamma)$, and the overall minimizing $\mathcal{B}_{\lambda(w)}$ can be found among the $\{\mathcal{B}_\lambda\}$ as stated in the theorem. □

EXAMPLE (Normal distributions). Suppose (T_i, H_i) are i.i.d. random variables, $T_i | H_i \sim (1-H_i)\cdot N(0,1) + H_i \cdot N(2,1)$, and $H_i \sim \mathrm{Bernoulli}(0.2)$. Also suppose we want to minimize

$$\tfrac{1}{3}\,\mathrm{pFDR}(\Gamma) + \tfrac{2}{3}\,\mathrm{pFNR}(\Gamma)$$

over all measurable Γ. Therefore, we have made the rate of nondiscoveries that are false two times as important as the rate that discoveries are false. By Theorem 6, we only have to consider significance regions of the form

$$\mathcal{B}_\lambda = \left\{ t : \frac{0.8\phi_{0,1}(t)}{0.8\phi_{0,1}(t) + 0.2\phi_{2,1}(t)} \le \lambda \right\},$$

where ϕ_{μ,σ^2} is the density of a $N(\mu, \sigma^2)$. By calculating $\lambda(2/3) = \arg\min_\lambda[1/3 \times \mathrm{pFDR}(\mathcal{B}_\lambda) + 2/3\,\mathrm{pFNR}(\mathcal{B}_\lambda)]$, we get $\lambda(2/3) = 0.193$, which implies $\mathcal{B}_{0.193} = \{T \ge 2.41\}$. Therefore $\inf_\Gamma 1/3\,\mathrm{pFDR}(\Gamma) + 2/3\,\mathrm{pFNR}(\Gamma) = 0.123$ and this occurs

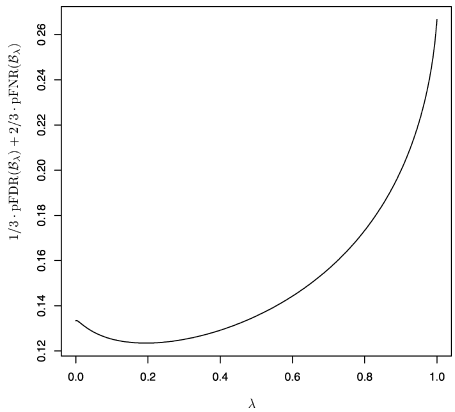

FIG. 3. *A plot of* $1/3 \cdot \mathrm{pFDR}(\mathcal{B}_\lambda) + 2/3 \cdot \mathrm{pFNR}(\mathcal{B}_\lambda)$ *as a function of* λ.

at $\Gamma = \mathcal{B}_{0.193} = \{T \geq 2.41\}$. Figure 3 shows $1/3\,\mathrm{pFDR}(\mathcal{B}_\lambda) + 2/3\,\mathrm{pFNR}(\mathcal{B}_\lambda)$ as a function of λ.

Since it will tend to be the case that $\pi_0 \gg \pi_1$, one may also wish to find Γ to minimize

$$(1-w) \cdot \frac{\mathrm{pFDR}(\Gamma)}{\pi_0} + w \cdot \frac{\mathrm{pFNR}(\Gamma)}{\pi_1}.$$

The minimizing set can also be found among the $\{\mathcal{B}_\lambda\}$ using some $\lambda'(w)$ defined similarly to the above.

7. An application to DNA microarrays in a Bayesian framework. Here we consider the application of some of these ideas to an empirical Bayesian approach to detecting differentially expressed genes in DNA microarray experiments. In doing so, we discuss the advantages and disadvantages of reporting the q-value as a measure of significance for each gene as opposed to reporting the classical posterior probability.

A DNA microarray allows the simultaneous measurement of the expression levels of thousands of genes from a single biological sample [Brown and Botstein (1999)]. Efron, Tibshirani, Storey and Tusher (2001) consider a data set in which four microarrays are obtained from "untreated" human cells, and four from irradiated human cells. Therefore, for each of over 6,000 genes, there are eight independent measurements. A modified two-sample t-statistic is calculated for each gene. It is assumed that $T_i | H_i \sim (1 - H_i) \cdot F_0 + H_i \cdot F_1$, as has been assumed in this work. Moreover, null versions of the statistics are calculated. Using these null versions along with the observed statistics, a nonparametric estimate of

$$\Pr(H_i = 0 | T_i = t_i) = \frac{\pi_0 \cdot f_0(t_i)}{\pi_0 \cdot f_0(t_i) + \pi_1 \cdot f_1(t_i)}$$

is calculated, which we denote by $\widehat{\Pr}(H_i = 0|T_i = t_i)$. The sets $\{\mathcal{B}_\lambda\}$ can then be estimated by $\widehat{\mathcal{B}}_\lambda = \{t : \widehat{\Pr}(H = 0|T = t) \leq \lambda\}$.

Efron, Tibshirani, Storey and Tusher (2001) suggest thresholding genes for differential gene expression by the significance region $\widehat{\mathcal{B}}_{0.10}$, which is equivalent to calling gene i differentially expressed if $\widehat{\Pr}(H_i = 0|T_i = t_i) \leq 0.10$. The threshold is determined by $\widehat{\Pr}(H_i = 0|T_i = t_i)$, which is only a marginal statistic and does not take into account the multiple comparisons. Therefore, the 0.10 used in the threshold does not have a straightforward interpretation in terms of the error rate of the overall list of genes. Using an earlier version of this work [Storey (2001)], they note that by integrating over $\widehat{\mathcal{B}}_{0.10}$ with the density $\widehat{f}(\cdot|T \in \widehat{\mathcal{B}}_{0.10})$, they have estimated $\text{pFDR}(\widehat{\mathcal{B}}_{0.10}) = \Pr(H = 0|T \in \widehat{\mathcal{B}}_{0.10})$. This is a clear illustration of how the results in this paper can be used in a Bayesian setting. There is a further issue, however, which is how to assign a measure of significance to each gene. Efron, Tibshirani, Storey and Tusher (2001) suggest reporting $\widehat{\Pr}(H_i = 0|T_i = t_i)$ as a measure of significance for each gene and then reporting $\text{pFDR}(\widehat{\mathcal{B}}_\lambda)$ according to which threshold λ is used. They argue $\widehat{\Pr}(H_i = 0|T_i = t_i)$ should be reported because it gives local information about the significance of the gene.

In this paper, we have defined the q-value as a pFDR measure of significance for each statistic. For this particular problem,

$$q\text{-value}(t_i) = \Pr\bigl(H_i = 0|T_i \in \mathcal{B}_{\Pr(H_i=0|T_i=t_i)}\bigr)$$

and it can be estimated by $\widehat{\Pr}(H_i = 0|T_i \in \widehat{\mathcal{B}}_{\widehat{\Pr}(H_i=0|T_i=t_i)})$. Whereas $\Pr(H_i = 0|T_i = t_i)$ provides a measure of significance local to t_i, q-value(t_i) provides a measure of significance in the same sense that the p-value does: It takes into account the fact that if we call gene i significant, then we are also forced to call all genes with greater evidence of differential expression significant. Moreover, the q-value simultaneously takes into account the multiple comparisons because it is defined in terms of the pFDR.

It is clear that both $\Pr(H_i = 0|T_i = t_i)$ and q-value(t_i) are valuable measures to consider. Ideally, we would have both at our disposal. We make the case here, though, that if only one measure of significance is to be used, then it should be q-value(t_i). This follows by the fact that in a multiple comparisons situation such as this, one always has to worry about controlling the number of false positives in some way. The interpretation of how this is being accomplished in q-value(t_i) is clear, whereas it is not in $\Pr(H_i = 0|T_i = t_i)$. A nonparametric method for estimating the q-values has been proposed in Storey (2002a), and it has been shown in Storey, Taylor and Siegmund (2004) that they are simultaneously conservatively consistent under fairly mild assumptions, even under certain forms of dependence. Much stronger assumptions have to be made to show that the $\widehat{\Pr}(H_i = 0|T_i = t_i)$ estimated in Efron, Tibshirani, Storey and Tusher (2001) are consistent and robust against dependence. Moreover, one can utilize the q-values in either a frequentist or Bayesian framework.

8. Discussion. False discovery rates are useful for multiple hypothesis testing in certain settings. They are especially useful when one is testing many hypotheses and wishes to have a low frequency of false positives among all the rejected hypotheses. We studied the pFDR, an alternative to the FDR, showing several interesting statistical properties. It has a simple Bayesian interpretation when the tests are independent and follow a mixture distribution. This Bayesian interpretation yields insight into the pFDR quantity. Moreover, it gives a multiple testing measure that can be used by Bayesians or frequentists. We showed how Efron, Tibshirani, Storey and Tusher (2001) used the results from this work to connect their empirical Bayesian method to false discovery rates.

The q-value is a natural counterpart to the p-value, especially under the mixture model. Since the q-value is concerned with the probability of a null hypothesis given the statistic is significant, it is a multiple hypothesis testing quantity, whereas the p-value is a single hypothesis testing quantity. It is hoped that the estimated q-value will be reported with each statistic when one does multiple hypothesis testing using the pFDR. The q-value was also shown to have several interesting properties, including a special relationship to the p-value under certain assumptions.

The pFDR was shown to have a very simple form under the i.i.d. assumption. Therefore, this quantity is quite tractable in practice. Even when dependence exists, the pFDR comes quite close to the form under independence when the number of tests gets large, as long as the dependence is weak enough to satisfy the conditions of Theorem 4. We calculated one such example with normal random variables. The pFDR and pFNR can be interpreted in the context of classification theory. The Bayes rule can be used to minimize the Bayes error (which is a weighted sum of the pFDR and the pFNR) or it can be used to minimize the weighted average of the pFDR and pFNR.

Acknowledgments. Thanks to Brad Efron, Rob Tibshirani, Larry Wasserman, the referees and an Associate Editor for helpful ideas and comments.

REFERENCES

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 HOCHBERG, Y. (2000). On the adaptive control of the false discovery rate in multiple testing with independent statistics. *J. Educational and Behavioral Statistics* **25** 60–83.

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

BILLINGSLEY, P. (1968). *Convergence of Probability Measures*. Wiley, New York.

BROWN, P. O. and BOTSTEIN, D. (1999). Exploring the new world of the genome with DNA microarrays. *Nature Genetics* **21** 33–37.

CHERKASSKY, V. S. and MULIER, F. M. (1998). *Learning from Data*: *Concepts*, *Theory and Methods*. Wiley, New York.

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.

GENOVESE, C. and WASSERMAN, L. (2002a). Operating characteristics and extensions of the procedure. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **64** 499–517.

GENOVESE, C. and WASSERMAN, L. (2002b). False discovery rates. Technical report, Dept. Statistics, Carnegie Mellon Univ.

LEHMANN, E. L. (1986). *Testing Statistical Hypotheses*, 2nd ed. Wiley, New York.

MORTON, N. E. (1955). Sequential tests for the detection of linkage. *Amer. J. Human Genetics* **7** 277–318.

SARKAR, S. K. (2002). Some results on false discovery rate in stepwise multiple testing procedures. *Ann. Statist.* **30** 239–257.

SHAFFER, J. (1995). Multiple hypothesis testing: A review. *Annual Review of Psychology* **46** 561–584.

SIMES, R. J. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika* **73** 751–754.

STOREY, J. D. (2001). The positive false discovery rate: A Bayesian interpretation and the q-value. Technical Report 2001-12, Dept. Statistics, Stanford Univ.

STOREY, J. D. (2002a). A direct approach to false discovery rates. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **64** 479–498.

STOREY, J. D. (2002b). False discovery rates: Theory and applications to DNA microarrays. Ph.D. dissertation, Dept. Statistics, Stanford Univ.

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. R. Stat. Soc. Ser. B Stat. Methodol.* **66** 187–205.

WELLER, J. I., SONG, J. Z., HEYEN, D. W., LEWIN, H. A. and RON, M. (1998). A new approach to the problem of multiple comparisons in the genetic dissection of complex traits. *Genetics* **150** 1699–1706.

ZAYKIN, D. V., YOUNG, S. S. and WESTFALL, P. H. (2000). Using the false discovery rate approach in the genetic dissection of complex traits: A response to Weller et al. *Genetics* **154** 1917–1918.

DEPARTMENT OF BIOSTATISTICS
UNIVERSITY OF WASHINGTON
SEATTLE, WASHINGTON 98195-7232
E-MAIL: jstorey@u.washington.edu