THE FALSE DISCOVERY RATE: A BAYESIAN INTERPRETATION AND THE q-value

by

John D. Storey

Technical Report No. 2001-12
May 2001

Department of Statistics
STANFORD UNIVERSITY
Stanford, California 94305

THE FALSE DISCOVERY RATE: A BAYESIAN INTERPRETATION AND THE q-value

by

John D. Storey
Department of Statistics
Stanford University

Technical Report No. 2001-12
May 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

The False Discovery Rate: A Bayesian Interpretation and the q-value

John D. Storey
Department of Statistics
Stanford University
Stanford, CA 94305
jstorey@stat.stanford.edu

May, 2001

Abstract

With the growing abundance of large data sets, multiple comparison procedures continue to gain importance. For example, active areas such as wavelet analysis and genomics often require one to essentially test many hypotheses simultaneously. One multiple comparison procedure is the False Discovery Rate, which measures the expected proportion of false positives among all significant hypotheses. In this paper we investigate some statistical properties of the False Discovery Rate. A Bayesian interpretation is made, and some asymptotic results are presented. Also, a new quantity called the *q-value* is introduced, which is the False Discovery Rate 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. 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 positive out of all hypotheses tested. The most familiar example of this is the Bonferroni method. If there are n hypothesis tests, each test is controlled so that the probability of a false positive is less than or equal to α/n for some chosen value of α. It then follows that the overall FWER is less than or equal to α. Many more methods have been introduced that improve upon the Bonferroni method in that the FWER is still controlled at level α, but the power for each test is increased. Shaffer (1995) provides a review of many of these methods.

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 possibility of making one or more false positive. The FWER offers an extremely strict criterion which is not always appropriate. We have seen a recent increase in the size of data sets available to statisticians. It is now often up to the statistician to find as many interesting features in a data set as possible rather than testing a very specific hypothesis on one item. For example, one is increasingly 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 rejected, rather than making one or more false positive.

1.1 The False Discovery Rate

Benjamini and Hochberg (1995) introduce a new multiple hypothesis testing error measure called the False Discovery Rate (FDR). If V is the number of false positives and R is the total number of null hypotheses rejected, then the FDR is defined to be $\mathbf{E}(V/R)$. When $R = 0$ the ratio V/R is set to zero. Benjamini and Hochberg (1995) present a p-value step-up method that allows one to control the FDR at a desired level when the hypothesis tests are independent. Suppose that the p-values resulting from the n tests

are ordered such that $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(n)}$. If we find

$$\hat{k} = \text{argmax}_{1 \leq k \leq n}\{k : p_{(k)} \leq \alpha \cdot k/n\},$$

then rejecting the hypotheses corresponding to $p_{(1)}, \ldots, p_{(\hat{k})}$ provides $\mathbf{E}(V/R) \leq \alpha$. (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 $\mathbf{E}(V/R) \leq \alpha$.) The FDR offers a less stringent control over Type I errors than the FWER, and is therefore usually more powerful, as is shown in their simulations.

1.2 Applications of the False Discovery Rate

Two examples of fields where the FDR has proven to be useful are wavelet analysis and genomics. Abramovich and Benjamini (1996) and Abramovich et al. (2000) both use the FDR for choosing a subset of wavelet coefficients out of thousands. It is impractical to guard against having one or more inappropriate choice in this situation. Therefore, the FDR provides a useful measure for making sure that most of the chosen coefficients are truly non-zero. Tusher et al. (2001) apply the FDR to a problem in DNA microarrays. DNA microarrays are a new biotechnology that allow measurement of the expression levels of thousands of genes simultaneously. Tusher et al. (2001) analyze an experiment in which the expression levels of several thousand human genes are compared between regular cells and cells treated with radiation. Essentially a two sample t-test is performed on each gene, and a permutation method is employed in order to estimate the FDR. In this case, it is overly conservative to worry about making one or even a few false positives. Over 6000 genes were tested with the hope of identifying many genes that should have shown differential gene expression between the treatment and control groups. Therefore, the FDR is also very appropriate for this kind of problem. Application of the FDR is not limited to these two areas. Benjamini and Hochberg (1995) explain very thoroughly when this method can be appropriately used. Also, a literature search will provide ample evidence that the FDR is of interest to researchers in a wide variety of fields.

1.3 Understanding the False Discovery Rate

The aim of this paper is to understand the FDR in terms of familiar statistical ideas. Benjamini and Hochberg (1995) give nice heuristic arguments as to why the FDR is a useful method. Their mathematical analysis is

limited to proving that the step-up p-value method provides control of the FDR. Also, several other papers have introduced other p-value based methods (e.g., Benjamini and Liu (1999)). Recently, Genovese and Wasserman (2001) investigated the "operating characteristics" of the FDR procedure. They study Benjamini and Hochberg's data dependent scheme for choosing a p-value cut-off. In this paper, we will be more concerned about studying the actual $\mathbf{E}[V/R]$ quantity. This will include an investigation into the definition and interpretation of the FDR, as well as some of the basic statistical properties of the quantity $\mathbf{E}[V/R]$. For example, is it appropriate to use the p-value to control and/or determine the FDR? Recall that the p-value is a quantity derived from Type I error, whereas the FDR also clearly involves the power.

The FDR is written as an expectation of a continuous quantity V/R, whereas the FWER is the probability of a particular event. Since the FDR averages over a random quantity, perhaps it has a Bayesian interpretation. Efron et al. (2001) attempt to solve the same DNA microarray problem as that of Tusher et al (2001). Efron et al. (2001) choose significant genes based on the posterior probability there is an effect given the data. Is there a connection between this posterior probability and the FDR reported in Tusher et al. (2001)?

We will show that there is a very direct connection between the frequentist FDR and this posterior probability. We will also show in what ways the p-value is related or unrelated to the FDR. Also, we will investigate the properties of the FDR when the tests are dependent. This paper is organized as follows. Section 2 discusses two possible definitions of the FDR, and proposes a new definition. Section 3 gives a Bayesian interpretation of the FDR when the hypothesis tests are independent. Section 4 introduces the q-value, the FDR analogue of the p-value, in the context of performing n identical hypothesis tests. Section 5 extends the definition of the q-value to an arbitrary set of hypothesis tests. Section 6 investigates the FDR when the tests are dependent and its relationship to the Bayesian interpretation, giving some asymptotic results and numerical examples. Section 8 provides remarks and proofs for all technical results.

2 Defining the False Discovery Rate

2.1 Two Possible Definitions of the FDR

Benjamini and Hochberg (1995) initially define the FDR to be

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

where V is the number of false positives and R is the total number of rejected null hypotheses for some rejection scheme. A problem arises when $R = 0$, so their final definition is

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

This quantity can be controlled by a p-value step-up procedure. From the discussion of the FDR in Benjamini and Hochberg (1995), one can easily become confused and think they are referring to the FDR as being the quantity

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

Definition (2) of the FDR can be written in words as "the rate that false discoveries occur", whereas definition (3) can be written as "the rate that discoveries are false." One has to ask which definition is more useful in practice. When a scientist is in a multiple hypothesis testing situation calling things significant by a certain criterion, s/he is probably not interested in cases where nothing is significant nor in controlling a quantity that involves cases where nothing is found.

An example where confusion between definitions (2) and (3) can be dangerous is the following. One can use the Benjamini and Hochberg (1995) procedure to guarantee $\mathbf{E}[V/R|R > 0]\mathbf{P}(R > 0) \leq 0.1$. But what if $\mathbf{P}(R > 0) = 0.5$? Then we have actually only controlled $\mathbf{E}[V/R|R > 0] \leq 0.2$, a quantity twice as large! One may suppose this example is hypothetical, but this exact confusion arises in Weller et al. (1998). Zaykin et al. (1998) show that the results of Weller et al. (1998) can be very misleading if definitions (2) and (3) are confused. Also, Shaffer (1995) states that the inclusion of $\mathbf{P}(R > 0)$ into the definition of the FDR is unsatisfying.

Why not avoid this confusion and control the quantity $\mathbf{E}[V/R|R > 0]$ instead? Benjamini and Hochberg (1995) point out that if all null hypotheses are true then $\mathbf{E}[V/R|R > 0] = 1$, so that this quantity cannot be controlled

in the traditional p-value based framework. Therefore, they choose to work with definition (2) even though much of their discussion of the FDR seems to point to quantity (3). One could argue that the $FDR = 1$ when all null hypotheses are true, especially when we want to measure "the rate that discoveries are false." Therefore, definition (2) can be a bit dubious from a practical viewpoint.

Given this discussion, it seems appropriate to consider a different definition of the FDR than definition (2). We will propose a new definition, not only because of the arguments given above, but because our definition shows some nice statistical properties, as will be seen. Hopefully, it will be shown that the new definition of the FDR is intuitively pleasing as well as mathematically tractable. To this end, we propose the following alternative definition of the FDR.

Definition 1 *We define the False Discovery Rate* – the rate at which discoveries are false – *to be:*

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

Multiple comparison procedures were developed before computers were readily available for use in statistics. Therefore, they had to be designed in terms of quantities that were able to be gathered from tables, such as tables of p-values for the normal or t distributions. Traditional p-value adjustment methods provide a bound on the Family Wise Error Rate (FWER). Benjamini and Hochberg (1995) introduce the FDR in this traditional context. The end product of these sequential p-value methods is an estimate $\hat{k}$ so that p-values $p_{(1)}, \ldots, p_{(\hat{k})}$ are rejected (where $p_{(1)}, \ldots, p_{(n)}$ are the ordered p-values). In other words, the acceptable error rate is set, and then the rejection region is estimated so that *on average* we have FDR $\leq \alpha$ (or FWER $\leq \alpha$). In the traditional statistical terms, we say that this procedure is conservatively biased. But if the estimate $\hat{k}$ has a high variance, then the given procedure would not be all that good on a case by case basis.

Providing cryptic sequential p-value methods to control the FDR or FWER essentially involves estimating the rejection region. This procedure makes it nearly impossible to estimate the reliability of any given $\hat{k}$ estimate. Therefore Storey (2001, in preparation) suggests a new approach to multiple hypothesis testing. Instead of fixing the error rate and then estimating the rejection region, he suggests fixing the rejection region and estimating the error rate. This allows one to take much more advantage of the data, as will be seen next. Also, one has all the power of point estimation to apply to

this procedure. From this approach one can use definition (2) or (3) of the FDR, so definition (3) of the FDR is tractable from this viewpoint.

2.2 Estimating $\mathbf{E}[V/R|R > 0]$

Storey (2001, in preparation) presents an accurate estimate of definition (3) that converges almost surely to a tight upper bound of the actual value. Moreover, for reasonable rejection regions, the estimate is conservatively biased. Our presentation of this is intended to give the reader an idea as to how there are different approaches one could take with respect to the FDR than using a sequential p-value method. Also, this shows that even though $\mathbf{E}[V/R|R > 0]$ is intractable with respect to a sequential p-value method, it can be easily handled by using a different approach.

We assume that there are n independent hypothesis tests. It will be shown in Section 3 that under independence

$$FDR(\gamma) = \frac{\gamma \cdot n_0/n}{\mathbf{P}(p \leq \gamma)} \tag{5}$$

when we reject all p-values less than or equal to γ. $\mathbf{P}(p \leq \gamma)$ is the probability *any* p-value is less than or equal to γ, so it is a mixture of n_0 null p-values and $n - n_0$ alternative p-values. Since n_0 of the p-values are 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 n_0 is

$$\hat{n}_0 = \frac{\#\{p_i \geq \beta\}}{1 - \beta} \tag{6}$$

for some well chosen β. It is shown in Storey (2001, in preparation) that taking β to be the median of the observed p-values works well. If n is large this is basically equivalent to setting $\beta = 0.5$. (Efron et al. (2001) take a similar approach to estimate n_0/n.) Also let

$$\hat{r} = \#\{i : p_{(i)} \leq \gamma\}. \tag{7}$$

In other words, we count how many observed p-values are less than or equal to γ.

We form our estimate of $FDR(\gamma)$ to be

$$\widehat{FDR}(\gamma) = \frac{\gamma \cdot \hat{n}_0}{\hat{r}} \tag{8}$$

Of course we would impose that $\widehat{FDR}(\gamma) \leq 1$ in practice. If we assume that the frequency of null hypotheses is maintained to be n_0/n in the asymptotic sense, then it can be shown that

$$\lim_{n\to\infty} \widehat{FDR}(\gamma) \geq \mathbf{E}\left[\frac{V(\gamma)}{R(\gamma)}\middle| R(\gamma) > 0\right] \text{ with probability 1.} \tag{9}$$

Therefore one can provide a well behaved point estimate of $FDR(\gamma)$ for *any* chosen γ. Also, if a confidence interval is desired, the observed p-values can be bootstrapped and a confidence interval can be formed with the $\widehat{FDR}(\gamma)*$ in the standard way (Efron and Tibshirani 1993). We will not investigate this approach to multiple hypothesis testing any further – a full treatment is given in Storey (2001, in preparation) and in Storey and Tibshirani (2001, in preparation). However, we want to make the point that it is possible to accurately estimate the FDR in many situations. Moreover, the estimate is not forced to be overly conservative, nor is the p-value cut-off forced by some predetermined method. Finally, it should be mentioned that this approach has been taken before. See Yekutieli and Benjamini (1999) and Tusher et al. (2001) for two examples.

In the remainder of this paper, the results are presented using our definition of the FDR. If one prefers the former definition, it will be clear how to convert the results, although it will result in less tractable formulas and ideas. Hopefully it will become clear that our definition is a more natural definition both in a practical sense and in a theoretical sense, not only from the arguments made above, but also from the way in which our definition of the FDR can be easily understood through familiar statistical ideas. Whatever one's opinion may be, it is clear that the quantity of interest in either definition of the FDR is $\mathbf{E}[V/R|R > 0]$.

3 A Bayesian Interpretation

In this section we present a Bayesian interpretation of the FDR. As it turns out, in many cases the FDR can be written in a very simple Bayesian form. Instead of basing the FDR on p-values as is done in Benjamini and Hochberg (1995), we will use the original statistics and fixed rejection regions. Rejecting hypotheses based on the p-values is a special case of this generalization. See Section 5 and Remark 1 in Section 8 for more about p-value based rejections. Also, our analysis will be from the perspective taken in Storey (2001) and Section 2.2. Therefore, we will treat the FDR as being derived from fixed rejection regions as opposed to data dependent rejection regions as in

Benjamini and Hochberg (1995). For a nice treatment of the FDR under data dependent rejection regions, see Genovese and Wasserman (2001).

Suppose we wish to perform n identical hypothesis tests of a simple null hypothesis versus a simple alternative hypothesis. First assume that statistics $X_1, \ldots, X_n$ are i.i.d., and we know their densities under the two hypotheses, f_0 and f_1. (Thus, the statistics are identically distributed in that they have the same distribution under either hypothesis.) For a given rejection region Γ, define the False Discovery Rate as we defined it in Section 2:

$$FDR(\Gamma) = \mathbf{E}\left[\frac{V(\Gamma)}{R(\Gamma)}\middle| R(\Gamma) > 0\right], \tag{10}$$

where $V(\Gamma)$ is the number of false positives and $R(\Gamma)$ is the total number of rejected hypotheses for rejection region Γ.

Let $H_i = 0$ when the null hypothesis is true and $H_i = 1$ when the alternative is true, $i = 1, \ldots, n$. Let $\mathbf{P}(H = 0)$ be the *a priori* probability that a hypothesis is true. To a frequentist $\mathbf{P}(H = 0)$ is the proportion of null hypotheses that are true. $\mathbf{P}(H = 0)$ can also parameterize the case where the H_i are i.i.d. Bernoulli random variables. To a Bayesian, $\mathbf{P}(H = 0)$ is the prior (subjective) probability that a null hypothesis is true. The following theorem allows a Bayesian interpretation of the FDR to be made.

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 $\mathbf{P}(H = 0)$. Then*

$$\mathbf{P}(H = 0 | X \in \Gamma) = FDR(\Gamma), \tag{11}$$

where $\mathbf{P}(H = 0)$ is the prior probability used in the calculation of $\mathbf{P}(H = 0 | X \in \Gamma)$.

From Theorem 1, a Bayesian interpretation of the FDR works as follows. First suppose we are in the Bayesian setting. The prior probability on $H = 0$ is $\mathbf{P}(H = 0)$ and on $H = 1$ is $\mathbf{P}(H = 1) = 1 - \mathbf{P}(H = 0)$. Our decision rule is to say $H = 0$ if $X \notin \Gamma$ and $H = 1$ if $X \in \Gamma$. As in the frequentist case, let V be the number of statistics X_i for which $H_i = 0$, yet we observe $X_i \in \Gamma$. Let R be the number of X_i for which $X_i \in \Gamma$. Thus, in Bayesian terms $FDR(\Gamma)$ is the loss function defined as the misclassification rate among statistics for which we classified $H = 1$. Note that when Γ is the entire sample space, say Ω, then

$$FDR(\Omega) = \mathbf{P}(H = 0 | X \in \Omega) = \mathbf{P}(H = 0). \tag{12}$$

$FDR(\Omega)$ is the (expected) proportion of hypotheses for which the null hypothesis is true. In the frequentist setting H is usually a fixed quantity. Therefore, in that sense the prior probability $\mathbf{P}(H = 0)$ is the proportion of hypotheses for which the null hypothesis is true.

A similar connection can be made when starting from a frequentist viewpoint. For the hypotheses $H_1, \ldots, H_n$, suppose that the null hypothesis holds for n_0 of them and the alternative hypothesis holds for $n - n_0$ of them. For a randomly selected *significant* statistic, the probability that it's a false positive is

$$\frac{0 \cdot (n - n_0)/n + \mathbf{P}(X \in \Gamma | H = 0) \cdot n_0/n}{\mathbf{P}(X \in \Gamma | H = 1) \cdot (n - n_0)/n + \mathbf{P}(X \in \Gamma | H = 0) \cdot n_0/n}. \tag{13}$$

It follows by a similar proof to Theorem 1 that

$$FDR(\Gamma) = \frac{\mathbf{P}(X \in \Gamma | H = 0)\mathbf{P}(H = 0)}{\mathbf{P}(X \in \Gamma)}, \tag{14}$$

where $\mathbf{P}(H = 0) = 1 - \mathbf{P}(H = 1) = n_0/n$. Even though we do not usually think of H as being a random variable in the frequentist setting, it follows by Bayes theorem that

$$FDR(\Gamma) = \mathbf{P}(H = 0 | X \in \Gamma). \tag{15}$$

Therefore, under the given assumptions, we have shown that when starting from the aforementioned Bayesian framework, the FDR emerges as the posterior probability a null hypothesis is true given its statistic is in the "rejection region" Γ. When starting from a frequentist framework, the FDR can be interpreted as the posterior probability a null hypothesis is true given its statistic is in the rejection region, with prior probabilities equaling the proportions of null and alternative hypotheses.

Remark: Note that without loss of generality we can assume the H_i are random. Suppose that the H_i are fixed and that n_0 out of n have a true null hypothesis. Then by setting $\mathbf{P}(H = 0) = n_0/n$, we get the same FDR treating the H_i as random. This is the case because the FDR is an expectation and the expected number of true null hypotheses is n_0/n whether we consider them to be random or not. Therefore, in order to be able to write the FDR in the frequentist or Bayesian form, we will treat the H_i as being random for the remainder of the paper.

In practice, the quantity $\mathbf{P}(H = 0|X \in \Gamma)$ has to be calculated. This can easily be done in a fully parametric Bayesian setting. In a frequentist or non-parametric setting, one or more of the quantities involved in calculating $\mathbf{P}(H = 0|X \in \Gamma)$ has to be estimated. See Storey (2001, in preparation) and Storey and Tibshirani (2001, in preparation) for more on this. Perhaps the best use of this interpretation is through an empirical Bayes analysis. In this situation, the prior probabilities on the hypotheses can be estimated from the data, either parametrically or non-parametrically as in Efron et al. (2001). Even from the Benjamini and Hochberg (1995) viewpoint, this Bayesian interpretation gives insight into the FDR quantity itself.

It is remarkable that the FDR can be written in such a simple form. $P(H = 0|X \in \Gamma)$ has an interesting intuitive interpretation: it is the probability of a null hypothesis given its statistic is in the rejection region, i.e., the frequency of null hypotheses with statistics falling into the rejection region. This latter phrase is a fairly accurate way to describe the FDR, so on one level it is a Bayesian quantity. On the other hand, the FDR can be calculated and interpreted completely from a frequentist point of view. Writing the FDR as $P(H = 0|X \in \Gamma)$ is very similar to the Type I error. One could call it a "posterior Bayesian Type I error." (See Morton (1955) for a very similar development of this concept in the context of genetic linkage analysis.) Whereas the FWER is very much frequentist, we have shown that the FDR 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.

Efron et al. (2001) also make the connection between a Bayesian posterior probability and the FDR. They define the "local false discovery rate" to be $\mathbf{P}(H = 0|X = x)$. This quantity is more precise than $\mathbf{P}(H = 0|X \in \Gamma)$ in that it gives the FDR for the rejection region that is a small neighborhood around $X = x$. Another measure that yields a more precise quantity than $FDR(\Gamma)$ is presented in the next section. Like the local false discovery rate, it gives a false positive rate to each observed statistic but in the context of the nested set of rejection regions and the FDR.

4 The q-value: A Definition and Some Properties

In this section, we introduce the FDR analogue of the p-value, which we call the *q-value*. Because of the connection made in the previous section, the q-value will be useful in both Bayesian and frequentist settings. It gives

the scientist a hypothesis testing error measure for each observed statistic with respect to the FDR. Again, assume that the statistics $X_1, \ldots, X_n$ have the same marginal distribution conditional on the null hypothesis, and conditional on the alternative hypothesis. As opposed to the previous section, the statistics may now be dependent in some arbitrary way. We introduce the q-value by first showing an example.

Example 4.1 *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 $p_i = \mathbf{P}(Z \geq z_i | H = 0)$. 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.

What is the FDR if we reject any statistic as extreme or more extreme than z_i among all n hypotheses? By Section 3, we can see that it is

$$q_i = \frac{n_0 \mathbf{P}(Z \geq z_i | H = 0)}{n_0 \mathbf{P}(Z \geq z_i | H = 0) + (n - n_0)\mathbf{P}(Z \geq z_i | H = 1)}. \tag{16}$$

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 1 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. Keep this simple example in mind as we formally introduce the q-value.

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

$$\text{p-value}(x) = \min_{\{\Gamma : x \in \Gamma\}} \mathbf{P}(X \in \Gamma | H = 0). \tag{17}$$

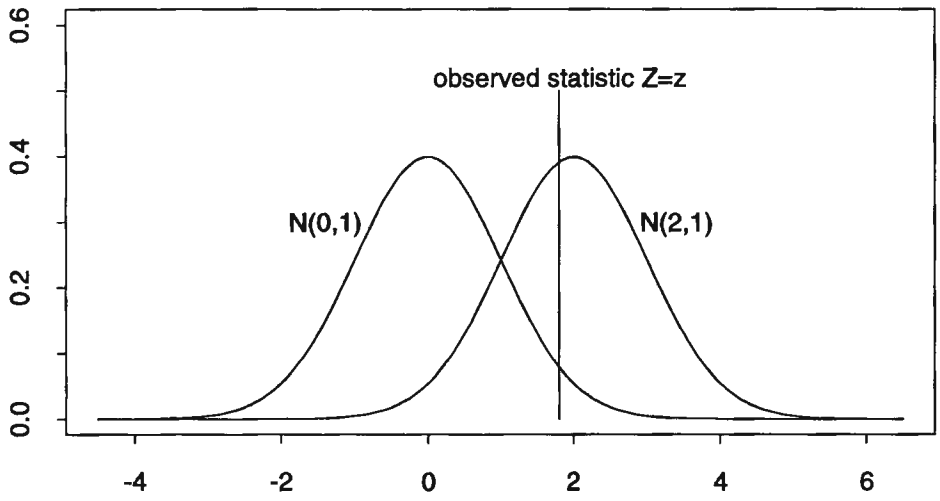

Figure 1: *A plot of the N(0,1) and N(2,1) densities. The vertical line denotes the observed statistic $Z = z$. The p-value and adjusted p-value can be calculated from the area under the N(0,1) density to the right of $Z = z$. The q-value is calculated using the area under both densities to the right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

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:*

$$q\text{-}value(x) = \min_{\{\Gamma:\ x \in \Gamma\}} FDR(\Gamma). \tag{18}$$

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. (Note this is assuming the statistics have the same marginal distribution under either hypothesis, and the rejection region is the same for every statistic. A more complicated definition would have to be made when the statistics have different rejection regions – see the next section for one such definition.)

When $n = 1$ or when the statistics are independent, we have that

$$\text{q-value}(x) = \min_{\{\Gamma:\ x \in \Gamma\}} \mathbf{P}(H = 0 | X \in \Gamma). \tag{19}$$

Therefore, the q-value is a Bayesian version of the p-value, say a "posterior Bayesian p-value", when we choose to regard the FDR as a Bayesian statement. Also, note that when the independence assumptions are met

$$\begin{aligned} &\text{argmin}_{\{\Gamma:\ x \in \Gamma\}} FDR(\Gamma) \\ &= \text{argmin}_{\{\Gamma:\ x \in \Gamma\}} \frac{\mathbf{P}(X \in \Gamma | H=0)\mathbf{P}(H=0)}{\mathbf{P}(X \in \Gamma | H=1)\mathbf{P}(H=1) + \mathbf{P}(X \in \Gamma | H=0)\mathbf{P}(H=0)} \qquad (20) \\ &= \text{argmin}_{\{\Gamma:\ x \in \Gamma\}} \frac{\mathbf{P}(X \in \Gamma | H=0)}{\mathbf{P}(X \in \Gamma | H=1)}. \end{aligned}$$

Therefore, in the case of independence, the q-value of a statistic minimizes the ratio of the Type I error to the power over all rejection regions that contain the statistic. This makes sense because the FDR is not as concerned with making false positives as with how frequent the false positives occur in relation to true positives.

One can understand this last observation in terms of a plot of power versus Type I error for a given set of rejection regions. (We will call this a power-Type I error plot.) Assume that this nested set of rejection regions produces a one-to-one map between Type I error and power so that we can write $\alpha = \mathbf{P}(X \in \Gamma | H = 0)$ and $g(\alpha) = \mathbf{P}(X \in \Gamma | H = 1)$ for some one-to-one function g. It can be shown through simple calculus that $g(\alpha)/\alpha$ is maximized at $\alpha = g(\alpha)/g'(\alpha)$, where g is also concave (down). Therefore we can maximize $g(\alpha)/\alpha$ 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 $g(\alpha)/\alpha$ is maximized.

See Figure 2 for a picture of this maximization. The left panel has a strictly concave $g(\alpha)$. Therefore, the ratio of power to Type I error decreases as $\alpha \to 0$. In other words, as the rejection regions get smaller, the ratio of power to Type I error gets larger. In view of our previous observation about the q-value, we can conclude that the Γ that minimizes $FDR(\Gamma)$ and $\mathbf{P}(X \in \Gamma | H = 0)/\mathbf{P}(X \in \Gamma | H = 1)$ also minimizes $\mathbf{P}(X \in \Gamma | H = 0)$. This follows since we would take the rejection region with the smallest $\alpha = \mathbf{P}(X \in \Gamma | H = 0)$ in order to maximize $g(\alpha)/\alpha$. Therefore, when the power-Type I error curve is concave, the same rejection region is used to minimize all three quantities.

This implies that the q-value of a statistic is based on the same rejection region as the p-value, as long as g is concave. We get concavity whenever the

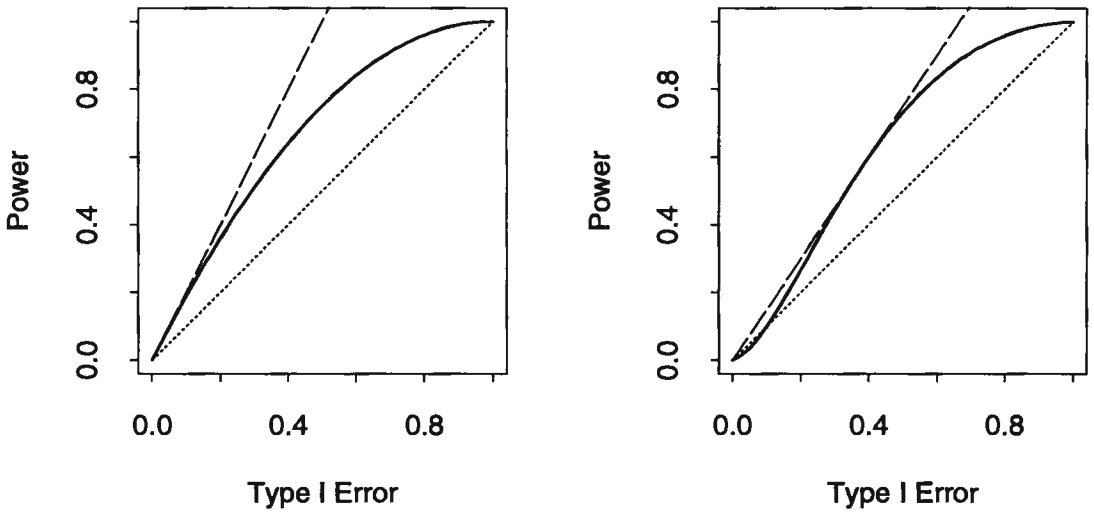

Figure 2: *A plot of power versus Type I error rate for two hypothetical sets of rejection regions. The solid line is power as a function of Type I error; the dotted line is the identity function; the dashed line is the line from the origin tangent to the power function.*

rejection regions are formed from a function that is monotone with respect to the likelihood ratio f_1/f_0. The right panel of Figure 2 shows an example where g is not concave. The rejection region that determines the q-value is the one which corresponds to the $(\alpha, g(\alpha))$ point that the line from the origin intersects. No similar connection can be made with the p-value under this kind of curve. Of course, one would hope to avoid this curve since over portions of the rejection regions we are better off doing a randomized test, but one could end up in an equivalent situation if certain forms of dependence existed between the hypotheses.

To conclude this section, we will show an example of how the q-value can be used in a Bayesian or empirical Bayes setting.

Example 4.2 *Detecting differential gene expression in DNA microarrays*

In Efron et al. (2000), scores $Z_1, \ldots, Z_n$ are calculated for n genes. The scores are similar to a two-sample t-statistic since they are testing for a difference in gene expression between treatment and control cells. The densities $f(z)$ and $f(z|H = 0)$ are estimated from the data as well as $\mathbf{P}(H = 0)$. For simplicity, we will suppose that these quantities are known. The rejection

regions used are $\{\Gamma_\alpha\}$ where

$$\Gamma_\alpha = \{z : \mathbf{P}(H = 0 \mid Z = z) \leq \alpha\}. \tag{21}$$

If the Z_i are treated as i.i.d., we see that

$$FDR(\Gamma_\alpha) = \mathbf{P}(H = 0 \mid Z \in \Gamma_\alpha). \tag{22}$$

We can also calculate the q-value of $Z = z$ as

$$\text{q-value}(z) = \mathbf{P}(H = 0 \mid Z \in \Gamma_{\mathbf{P}(H=0|Z=z)}). \tag{23}$$

Even if there is "loose" dependence between the scores, we will see in Section 6 that the calculations done for the FDR and q-value under the assumption of independence still approximately hold since $n = 6810$ in this example. For each gene, one is able to report its q-value, which gives its relative strength of being unaffected (the smaller, the better) compared to the other genes. □

We can see from the two examples in this section that the q-value can be useful in either a frequentist or Bayesian context. This is not the case for the adjusted p-value. Therefore, in an empirical Bayes setting, such as Efron et al. (2001), the q-value is a multiple testing measure that is interpretable in both schools of thought.

For a given rejection scheme, one can control the FDR at a certain level. However, something more can be said about the FDR for each observed statistic. The q-value accomplishes this quite nicely. It gives the FDR when rejecting statistics "as extreme or more extreme" than what was observed. This is done in the context of the original set of nested rejection regions, just like the p-value.

5 The q-value for Arbitrary Hypothesis Tests

For completeness, we will introduce a definition of the q-value that can be applied to any set of hypothesis tests. Suppose that for hypothesis tests $H_1, \ldots, H_n$ we have sets of nested rejection regions $\{\Gamma_1\}, \ldots, \{\Gamma_n\}$ and statistics $X_1, \ldots, X_n$, respectively. We need a mapping of the statistics based on the rejection regions so that each statistic can be compared, and therefore rejected in a uniform fashion. The most obvious choice for this mapping is the p-value for each test, say, $p_1, \ldots, p_n$. Therefore, any $X_i = x_i$ yields a p-value p_i. From this we can define the q-value in the following way.

Definition 3 *Suppose $H_1, \ldots, H_n$ are n arbitrary hypothesis tests with nested sets of rejection regions $\{\Gamma_1\}, \ldots, \{\Gamma_n\}$ and statistics $X_1, \ldots, X_n$, respectively. Define the q-value of $X_i = x_i$ with p-value p_i to be*

$$q\text{-}value(x_i) = q\text{-}value(p_i) = \min_{\{\alpha \geq p_i\}} FDR(\{p : p \leq \alpha\}), \tag{24}$$

where $FDR(\{p : p \leq \alpha\})$ is the FDR obtained when we reject all hypotheses with p-value less than or equal to α.

A natural question to ask is whether this definition is equivalent to the first definition of the q-value when the n hypothesis tests are identical. The following lemma and theorem show that the two definitions coincide. The theorem also shows when we can write the "p-value based" q-value in a more natural form.

Lemma 1 *For n identical hypothesis tests $FDR_X(\Gamma) = FDR_p(p_\Gamma)$, where $FDR_X(\Gamma)$ is the FDR based on the original statistics $X_1, \ldots, X_n$ and rejection region Γ, $FDR_p(p_\Gamma)$ is the FDR based on the p-values of $X_1, \ldots, X_n$, and $p_\Gamma = \mathbf{P}(X \in \Gamma | H = 0)$.*

Using this lemma and the discussion in the previous section, we can prove the next theorem.

Theorem 2 *For n identical hypothesis tests with nested set of rejection regions $\{\Gamma\}$, the two definitions of the q-value coincide. Also, it follows that when the tests are independent*

$$q\text{-}value(x) = FDR(\{p : p \leq p\text{-}value(x)\}) \tag{25}$$

if and only if the power-Type I error function is concave.

Therefore the more general definition of the q-value is a natural extension of the one presented in the previous section. The second statement in Theorem 2 shows that the most intuitive p-value based definition of the q-value is true only when we have independence and a concave power-Type I error function. We conclude this section with an example.

Example 5.1 *Arbitrary hypothesis tests*

We have statistics $X_1, \ldots, X_n$ corresponding to hypothesis tests $H_1, \ldots H_n$, and each test has its own set of nested rejection regions. If we calculate the

p-values $p_1, \ldots, p_n$ for each observed statistic $X_1 = x_1, \ldots, X_n = x_n$, then the q-value of p_i is

$$\text{(26)} \qquad \text{q-value}(p_i) = \min_{\{\alpha \geq p_i\}} FDR(\{\text{reject all } p_j \text{ such that } p_j \leq \alpha\})$$

Under dependence, we can numerically calculate the FDR from the observed data based on methods in Storey and Tibshirani (2001, in preparation) or in Yekutieli and Benjamini (1999).

If we have independence, the q-values can be calculated by using the method of Storey (2001, in preparation). Suppose that the function $r(i)$ gives the ascending ranking of the p_i. Then rejecting all p_j less than or equal to p_i gives an estimate of the FDR as $\hat{n}_0 \cdot p_i / r(i)$, where $\hat{n}_0$ is defined in Storey (2001, in preparation). Therefore, the q-value of p_i (and x_i) under our definition of the FDR is

$$\text{(27)} \qquad \text{q-value}(p_i) = \frac{\hat{n}_0 \cdot p_i}{r(i)}.$$

We can use $1 - (1 - p_i)^n$ as an estimate of the extra factor in the Benjamini and Hochberg (1995) FDR to get $\text{q-value}(p_i) = \hat{n}_0 \cdot p_i / r(i) \cdot [1 - (1 - p_i)^n]$ under their FDR. □

6 Dependence

Once again, in this section we assume we are testing n identical hypotheses $H_1, \ldots, H_n$ based on statistics $X_1, \ldots, X_n$ with rejection region Γ. When the statistics have the same marginal distribution (conditional and unconditional on H) but are dependent, we may write

$$\text{(28)} \quad FDR(\Gamma) = \sum_{k=1}^{n} \mathbf{P}\left(H_1 = 0 \,\middle|\, \begin{array}{c} X_1, \ldots, X_k \in \Gamma \\ X_{k+1}, \ldots, X_n \notin \Gamma \end{array} \right) \cdot \mathbf{P}(R = k | R > 0).$$

From this expression it can be seen that Theorem 1 does not hold in general under dependence. It is interesting to determine when Theorem 1 holds approximately, or rather asymptotically. Two theorems are presented in this section that address this issue. Also, some numerical examples are given to demonstrate the results.

To make the results as general as possible, we will define exactly how the statistics $X_1, \ldots, X_n$ are generated for each n. We want the statistics to be exchangeable for each fixed n (so that, for example, each hypothesis test uses the same rejection region). We do not want to force exchangeability in

the asymptotic sense because this limits the classes of random variables we can explore. Therefore, for each n we assume the experimenter obtains the statistics $X_1, X_2, \ldots, X_n$ in the following way:

- According to the a priori probabilities, $\mathbf{P}(H = 0)$ and $\mathbf{P}(H = 1)$, N_0 statistics come from the null distribution and N_1 from the alternative distribution, where $N_0 + N_1 = n$.
- Two independent sequences of random variables exist, $Y_1, Y_2, \ldots$ and $Z_1, Z_2, \ldots$, which follow the null and alternative distributions, respectively.
- These sequences are strongly stationary in the sense that $Y_1, Y_2, \ldots, Y_k$ is equal in distribution to $Y_{l+1}, Y_{l+2}, \ldots, Y_{l+k}$ for any $l \geq 0$ and $k \geq 1$. The same holds for $Z_1, Z_2, \ldots$. Note that this implies the Y_i have common marginal density f_0, and the Z_i have common marginal density f_1.
- For some random permutation of the indices $\sigma(1), \ldots, \sigma(n)$, we have $X_{\sigma(1)} = Y_1, \ldots, X_{\sigma(N_0)} = Y_{N_0}$, and $X_{\sigma(N_0+1)} = Z_1, \ldots, X_{\sigma(n)} = Z_{N_1}$.

Since we are proving results for the ratio V/R, it is valid that the permutation is for each fixed n. Another way to explain this last step is that the experimenter forms his or her statistics without regard to the indices (i.e. there is no natural ordering of the data). If a natural ordering of the data exists, the following two theorems can easily be adjusted for that particular situation.

The first result of this section shows that when $Y_1, Y_2, \ldots$ and $Z_1, Z_2, \ldots$ are ergodic, we get convergence of the FDR as $n \to \infty$ to the simple Bayesian form we obtained under independence. From a frequentist perspective, it shows that under the same conditions, we get convergence to the FDR one would obtain under independence.

Theorem 3 *Suppose that the statistics $X_1, X_2, \ldots$ are generated from $Y_1, Y_2, \ldots$ and $Z_1, Z_2, \ldots$, which are ergodic sequences of random variables under the shift operator. Then for any Lebesgue measurable rejection region Γ,*

$$lim_{n\to\infty} FDR_n(\Gamma) = \mathbf{P}(H = 0 | X \in \Gamma), \tag{29}$$

where $FDR_n(\Gamma)$ is the FDR resulting from the first n statistics.

In general, Theorem 3 says that if the statistics under either hypothesis are not "too correlated" then convergence to the simple Bayesian form

(or independence form) of the FDR occurs. Roughly speaking, not "too correlated" would indicate that the dependence between two Y_i (or two Z_i) decreases at an appropriate rate as the distance between their indices increases. Also, note that Theorem 3 allows many of the properties shown for the q-value under independence to hold approximately under ergodic dependence.

Example 6.1 *Locally correlated alternative statistics*

As a numerical example to illustrate the result of Theorem 3, consider the following situation. Suppose $Y_i \sim N(0,1)$ and $Z_i \sim N(\mu,1)$. The Y_i are independent, but $\mathbf{Cov}(Z_i, Z_{i+k}) = \rho$ where $0 \leq \rho \leq 1$ for $k = 1, 2, \ldots, 9$ and $i = 1, 11, 21, \ldots$, and zero covariance otherwise. In other words the Z_i have correlation ρ in groups of 10.

Suppose we take $\Gamma = [c, \infty)$ for some c. By Theorem 3, $FDR_n([c,\infty)) \rightarrow \mathbf{P}(H = 0 | X \in [c,\infty))$. Table 1 shows a comparison of the limiting FDR ($n = \infty$) compared to the FDR at a variety of finite values of n . The limiting FDR is calculated using the simple Bayesian form under independence. The finite cases are calculated by a Monte Carlo simulation (with 1000 iterations on `S-plus`). We use $\mu = 2$, $c = 2$ or $c = 3$, and $\mathbf{P}(H = 0) = 0.9$. It can be seen that there is quite good agreement between the limiting case and the finite cases, especially for large n. Most of differences between $n = 5000$ and $n = \infty$ are within Monte Carlo error. Convergence takes longer as ρ or c increase. This follows because the averaged sums of indicator variables used in the proof increase their variance as ρ or c increase. □

Even when there is strong correlation, we are able to write the limiting FDR in a closed form by Theorem 4 below. This does not have a simple Bayesian interpretation, but it does allow us to make convenient calculations, as is shown in the example that follows.

Theorem 4 *If the statistics $X_1, X_2, \ldots$ are generated from $Y_1, Y_2, \ldots$ and $Z_1, Z_2, \ldots$, which are strongly stationary sequences of random variables, then*

(30)

$$lim_{n\rightarrow\infty} FDR_n(\Gamma) = \mathbf{E}\left[\frac{W_0 \cdot \mathbf{P}(H=0)}{W_0 \cdot \mathbf{P}(H=0) + W_1 \cdot \mathbf{P}(H=1)} \middle| W_0 + W_1 > 0\right]$$

where W_0 and W_1 are random variables with support on $[0,1]$ defined in the proof. The limiting quantity is not equal to $\mathbf{P}(H = 0 | X \in \Gamma)$ in general.

Example 6.2 *Globally correlated alternative statistics*

Table 1: False Discovery Rate for Example 6.1

$c = 2$						
n	$\rho = 0$	$\rho = 0.10$	$\rho = 0.25$	$\rho = 0.50$	$\rho = 0.75$	$\rho = 1$
100	0.272	0.288	0.313	0.355	0.419	0.548
500	0.286	0.290	0.288	0.300	0.302	0.328
1000	0.289	0.291	0.291	0.293	0.299	0.304
5000	0.291	0.290	0.292	0.291	0.292	0.293
∞	0.291	0.291	0.291	0.291	0.291	0.291
$c = 3$						
n	$\rho = 0$	$\rho = 0.10$	$\rho = 0.25$	$\rho = 0.50$	$\rho = 0.75$	$\rho = 1$
100	0.0692	0.0893	0.0979	0.138	0.175	0.345
500	0.0720	0.0766	0.0745	0.0934	0.128	0.274
1000	0.0695	0.0760	0.0747	0.0810	0.0995	0.179
5000	0.0714	0.0721	0.0717	0.0725	0.0727	0.0743
∞	0.0711	0.0711	0.0711	0.0711	0.0711	0.0711

An example where Theorem 4 is useful is the following situation. We have that $Y_i \sim N(0,1)$ and $Z_i \sim N(\mu,1)$. The Y_i are independent, but $\mathbf{Cov}(Z_i, Z_j) = \rho$ where $0 \leq \rho \leq 1$ for $i \neq j$. Suppose we take $\Gamma = [c, \infty)$ for some c. By the proof of Theorem 3, $W_0 = \mathbf{P}(Y \geq c) = \mathbf{P}(X \in \Gamma | H = 0)$ with probability 1. It can be shown (see Section 8) that W_1 has c.d.f.

$$F(w) = 1 - \Phi\left(\frac{\Phi^{-1}(1-w)}{\sqrt{\rho/(1-\rho)}} - \frac{c-\mu}{\sqrt{\rho}}\right), \tag{31}$$

for $0 \leq w \leq 1$, where Φ is the c.d.f. for a $N(0,1)$ random variable.

In particular, when $\rho = 0$, $W_1 = \mathbf{P}(Z \geq c)$; when $\rho = 1$, $W_1 = 1(Z \geq c)$; and when $\rho = 1/2$ and $c = \mu$, $W_1 \sim \text{Unif}[0,1]$. Table 2 shows a comparison of the limiting FDR ($n = \infty$) values compared to the FDR at a variety of finite values of n. The limiting FDR is calculated by using the above c.d.f. to numerically approximate the integral in Theorem 4. The finite cases are calculated by a Monte Carlo simulation (again 1000 iterations in `S-plus`). We use $\mu = 2$, $c = 2$ or $c = 3$, and $\mathbf{P}(H = 0) = 0.9$ as before. It can be seen that there is quite good agreement between the limiting case and the finite cases, especially when n is large. The same trends in convergence exist as in the previous example, for the same reasons. □

Table 2: False Discovery Rate for Example 6.2

$c = 2$						
n	$\rho = 0$	$\rho = 0.10$	$\rho = 0.25$	$\rho = 0.50$	$\rho = 0.75$	$\rho = 1$
100	0.276	0.289	0.299	0.343	0.421	0.553
500	0.286	0.298	0.315	0.357	0.425	0.584
1000	0.288	0.298	0.319	0.362	0.431	0.583
5000	0.290	0.298	0.317	0.359	0.410	0.586
∞	0.291	0.301	0.319	0.363	0.425	0.585
$c = 3$						
n	$\rho = 0$	$\rho = 0.10$	$\rho = 0.25$	$\rho = 0.50$	$\rho = 0.75$	$\rho = 1$
100	0.0739	0.0869	0.103	0.125	0.162	0.419
500	0.0727	0.0901	0.125	0.197	0.343	0.697
1000	0.0694	0.0873	0.128	0.230	0.402	0.788
5000	0.0707	0.0895	0.137	0.266	0.425	0.845
∞	0.0711	0.0892	0.135	0.259	0.421	0.843

7 Discussion

The False Discovery Rate is a very useful error measure for multiple hypothesis testing. It is especially useful when one is testing many hypotheses and wishes to have a low frequency of false positives among all the rejected hypotheses. There has been one definition of the FDR considered until this paper. We proposed a second definition because it is most likely the quantity of interest to scientists, and the Benjamini and Hochberg (1995) definition can be quite misleading. The new definition of the FDR – "the rate that discoveries are false" – shows several interesting statistical properties. It has a simple Bayesian interpretation when the tests are independent. This Bayesian interpretation yields insight into the FDR quantity. Moreover, it gives a multiple testing measure that can be used by Bayesians or frequentists.

The q-value is a natural extension of the p-value. It is hoped that the q-value will be reported with each statistic when one does multiple hypothesis testing using the FDR. The q-value was also shown to have several interesting properties, including a special relationship to the p-value when the power-Type I error curve is concave.

Our definition of the FDR was shown to have a very simple form under independence. Therefore, this quantity is quite tractable in practice. Even when dependence exists, the FDR comes quite close to the form under inde-

pendence when the number of tests gets large as long as the dependence is ergodic. When the dependence is stronger than ergodic dependence, an explicit asymptotic form can nevertheless be found that is not too complicated. We calculated one such example with normal random variables.

We mostly focused on a simple versus simple hypothesis testing situation. It is possible to extend this to composite alternative hypotheses; see Genovese and Wasserman for the reasoning behind this.

The new definition of the FDR cannot be controlled with a p-value based method. This is unfortunate because it is clearly very useful in practice. Future work will show that this is not as much of a problem as it appears to be, and more powerful methods emerge by taking another approach to multiple hypothesis testing.

8 Remarks and Proofs

Remark 1:

One can view a p-value as a mapping from the statistic to a standardized value in the unit interval. The hypotheses are rejected based on the values obtained from this mapping. One could just as easily reject a hypothesis based on the original statistic. In non-parametric situations, converting to p-values may lead to a loss of information. For example, if one has to simulate the null distribution from the data, the p-values can have a limited number of values they can take, so that statistics cannot be rejected in a continuous fashion. Calculating the FDR directly with the original statistics circumvents this problem. In Efron et al. (2001), the data they work with has only 16 permutations available to estimate the null distribution. This puts the p-value on the 1/16 scale for doing over 6000 hypothesis tests. On average we would have to reject over 350 hypotheses with each increasing p-value. This is clearly undesirable, and it is preferable to work with the original, continuous statistics. Also, when the tests have an unknown dependence structure, the p-value based methods break down, so one is forced to estimate the FDR. Therefore, when dependence exists or in a non-parametric situation, one can choose to estimate quantity (2) or (3), whichever is more appropriate. See Storey and Tibshirani (2001, in preparation) for a treatment of these two cases.

Proof of Theorem 1:

First note that

$$FDR(\Gamma) = \mathbf{E}\left[\frac{V}{R}\middle| R>0\right] \tag{32}$$

$$= \sum_{k=1}^{n} \mathbf{E}\left[\frac{V}{R}\middle| R=k\right] \mathbf{P}(R=k|R>0) \tag{33}$$

$$= \sum_{k=1}^{n} \mathbf{E}\left[\frac{V}{k}\middle| R=k\right] \mathbf{P}(R=k|R>0). \tag{34}$$

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

$$\frac{\mathbf{P}(X\in\Gamma|H=0)\cdot\mathbf{P}(H=0)}{\mathbf{P}(X\in\Gamma)} = \mathbf{P}(H=0|X\in\Gamma). \tag{35}$$

Therefore,

$$FDR(\Gamma) = \sum_{k=1}^{n} \frac{k\cdot\mathbf{P}(H=0|X\in\Gamma)}{k} \mathbf{P}(R=k|R>0) \tag{36}$$

$$= \mathbf{P}(H=0|X\in\Gamma). \tag{37}$$

Proof of Lemma 1:

Let $p(x)$ be the p-value of $X=x$. Because the set of rejection regions is nested, it is trivial to show that $p(x)\leq p_\Gamma$ if and only if $x\in\Gamma$. This gives the equality.

Remark 2:

It trivially follows from Lemma 1 that for n independent hypothesis tests,

$$FDR_X(\Gamma_1,\ldots,\Gamma_n) = FDR_p(p_{\Gamma_1},\ldots,p_{\Gamma_n}). \tag{38}$$

$FDR_X(\Gamma_1,\ldots,\Gamma_n)$ is the FDR based on the original statistics $X_1,\ldots,X_n$ and rejection regions $\Gamma_1,\ldots,\Gamma_n$. $FDR_p(p_{\Gamma_1},\ldots,p_{\Gamma_n})$ is the FDR based on the p-values of $X_1,\ldots,X_n$ (with respect to the sets of nested rejection regions), and $p_{\Gamma_i}=\mathbf{P}(X_i\in\Gamma_i|H_i=0)$. By independence, the FDR essentially is a ratio of some subset of n Bernoulli random variables. Therefore, the proof follows by using the previous argument on each X_i and Γ_i.

Proof of Theorem 2:

For any $X = x$, let $\Gamma' = \text{argmin}_{\{\Gamma:\ x \in \Gamma\}} FDR_X(\Gamma)$. Then $x \in \Gamma'$ and thus $p(x) \leq p_{\Gamma'}$, where $p(x)$ is the p-value of $X = x$. By Lemma 1, $FDR_X(\Gamma) = FDR_p(p_\Gamma)$, so that $p_{\Gamma'} = \text{argmin}_{\{p_\Gamma : p(x) \leq p_\Gamma\}} FDR_p(p_\Gamma)$.

For the second statement, first suppose that g, the power-Type I error curve, is concave. For any $X = x$, let $\Gamma' = \text{argmin}_{\{\Gamma:\ x \in \Gamma\}} FDR_X(\Gamma)$. Then q-value$(x) = FDR(\Gamma')$. By concavity of the power-Type I error curve, $\Gamma' = \text{argmin}_{\{\Gamma:\ x \in \Gamma\}} \mathbf{P}(X \in \Gamma | H = 0)$, i.e., $p(x) = p_{\Gamma'} = \mathbf{P}(X \in \Gamma' | H = 0)$. By Lemma 1, $FDR_X(\Gamma') = FDR_p(\{p \leq p_{\Gamma'}\})$. Since the two definitions of q-value coincide in this case, we have the result.

Now suppose that q-value$(x) = FDR(\{p : p \leq p(x)\})$ for each x. Therefore, q-value(x) is an increasing function of $p(x)$. This implies $g(\alpha)/\alpha$ is a decreasing function of α. Since $g(0) = 0$ and $g(1) = 1$, it follows g is concave.

Proof of Theorem 3:

Let Γ be a Lebesgue measurable rejection region. Define for $i = 1, 2, \ldots$

$$I_i = \begin{cases} 0 & \text{if } Y_i \notin \Gamma \\ 1 & \text{if } Y_i \in \Gamma \end{cases}, \quad J_i = \begin{cases} 0 & \text{if } Z_i \notin \Gamma \\ 1 & \text{if } Z_i \in \Gamma \end{cases}. \tag{39}$$

Since $Y_1, Y_2, \ldots$ is ergodic and Γ is measurable, $I_1, I_2, \ldots$ is ergodic. Similarly, $J_1, J_2, \ldots$ is ergodic. Let N_0 be the number of i such that $H_i = 0$, and let N_1 be the number of i such that $H_i = 1$, where $N_0 + N_1 = n$. Then

$$\frac{V}{R} = \frac{\sum_{i=1}^{N_0} I_i}{\sum_{i=1}^{N_0} I_i + \sum_{j=1}^{N_1} J_j} \tag{40}$$

$$= \frac{\frac{N_0}{n} \frac{\sum_{i=1}^{N_0} I_i}{N_0}}{\frac{N_0}{n} \frac{\sum_{i=1}^{N_0} I_i}{N_0} + \frac{N_1}{n} \frac{\sum_{j=1}^{N_1} J_j}{N_1}}. \tag{41}$$

By independence of the H_i and the strong law of large numbers, as $n \to \infty$

$$N_0/n \xrightarrow{a.s.} \mathbf{P}(H = 0), \tag{42}$$

$$N_1/n \xrightarrow{a.s.} \mathbf{P}(H = 1). \tag{43}$$

By the ergodic theorem (e.g., Billingsley 1995) we have that as $n \to \infty$

$$\frac{\sum_{i=1}^{n} I_i}{n} \xrightarrow{a.s.} \mathbf{E}(I_i) = \mathbf{P}(X_i \in \Gamma | H = 0), \tag{44}$$

$$\frac{\sum_{i=1}^{n} J_i}{n} \xrightarrow{a.s.} \mathbf{E}(J_i) = \mathbf{P}(X_i \in \Gamma | H = 1). \tag{45}$$

Since $N_0 \xrightarrow{a.s.} \infty$ and $N_1 = n - N_0 \xrightarrow{a.s.} \infty$, it follows that as $n \to \infty$

$$\frac{\sum_{i=1}^{N_0} I_i}{N_0} \xrightarrow{a.s.} \mathbf{P}(X \in \Gamma | H = 0), \tag{46}$$

$$\frac{\sum_{j=1}^{N_1} J_j}{N_1} \xrightarrow{a.s.} \mathbf{P}(X \in \Gamma | H = 1). \tag{47}$$

Therefore, as $n \to \infty$

$$\frac{\frac{N_0}{n}\frac{\sum_{i=1}^{N_0} I_i}{N_0}}{\frac{N_0}{n}\frac{\sum_{i=1}^{N_0} I_i}{N_0} + \frac{N_1}{n}\frac{\sum_{j=1}^{N_1} J_j}{N_1}} \xrightarrow{a.s.} \mathbf{P}(H = 0 | X \in \Gamma). \tag{48}$$

It finally follows that

(49)

$$\lim_{n\to\infty} FDR_n(\Gamma) = \lim_{n\to\infty} \mathbf{E}_n \left[\frac{V}{R} \middle| R > 0 \right] = \lim_{n\to\infty} \mathbf{E}_n \left[\frac{V}{R} \right] = \mathbf{P}(H = 0 | X \in \Gamma),$$

where the second equality follows since $\lim_{n\to\infty} \mathbf{P}_n(R > 0) = 1$.

Proof of Theorem 4:

Using the definitions from the previous proof, we have $I_1, I_2, \ldots$ and $J_1, J_2, \ldots$ are independent, strongly stationary sequences of random variables. By using a similar argument as before we have

$$\frac{V}{R} = \frac{\frac{N_0}{n}\frac{\sum_{i=1}^{N_0} I_i}{N_0}}{\frac{N_0}{n}\frac{\sum_{i=1}^{N_0} I_i}{N_0} + \frac{N_1}{n}\frac{\sum_{i=1}^{N_1} J_i}{N_1}}. \tag{50}$$

Also, similarly as $n \to \infty$, $\frac{N_0}{n} \xrightarrow{a.s.} \mathbf{P}(H = 0)$ and $\frac{N_1}{n} \xrightarrow{a.s.} \mathbf{P}(H = 1)$.

By Theorem 24.1 of Billingsley (1995), a weaker ergodic theorem than the one used in the previous proof, it follows that

$$\lim_{n\to\infty} \frac{\sum_{i=1}^{n} I_i}{n} \overset{a.s.}{=} W_0, \tag{51}$$

$$\lim_{n\to\infty} \frac{\sum_{i=1}^{n} J_i}{n} \overset{a.s.}{=} W_1, \tag{52}$$

where W_0 and W_1 are integrable random variables such that $\mathbf{E}(W_0) = \mathbf{P}(X \in \Gamma | H = 0)$ and $\mathbf{E}(W_1) = \mathbf{P}(X \in \Gamma | H = 1)$. Both W_0 and W_1

have support on $[0,1]$. Since $N_0 \xrightarrow{a.s.} \infty$ and $N_1 = n - N_0 \xrightarrow{a.s.} \infty$, it follows that

$$\lim_{n\to\infty} \frac{\sum_{i=1}^{N_0} I_i}{N_0} \stackrel{a.s.}{=} W_0, \tag{53}$$

$$\lim_{n\to\infty} \frac{\sum_{i=1}^{N_1} J_i}{N_1} \stackrel{a.s.}{=} W_1. \tag{54}$$

Therefore, we can conclude that

$$\lim_{n\to\infty} \frac{V}{R} \stackrel{a.s.}{=} \frac{W_0 \cdot \mathbf{P}(H=0)}{W_0 \cdot \mathbf{P}(H=0) + W_1 \cdot \mathbf{P}(H=1)}. \tag{55}$$

By taking expectations we get

$$\lim_{n\to\infty} FDR_n(\Gamma) = \lim_{n\to\infty} \mathbf{E}_n \left[\frac{V}{R} \middle| R > 0 \right] \tag{56}$$

$$= \mathbf{E}\left[\frac{W_0 \cdot \mathbf{P}(H=0)}{W_0 \cdot \mathbf{P}(H=0) + W_1 \cdot \mathbf{P}(H=1)} \middle| W_0 + W_1 > 0 \right]. \tag{57}$$

The last quantity is not equal to $\mathbf{P}(H = 0 | X \in \Gamma)$ in general.

Remark: Note that this is a generalization of the previous proof. With ergodic sequences, W_0 and W_1 both have point masses on their expectations.

Calculating the c.d.f. of W_1 in Example 6.2:

We can write $Z_i = \mu + Y + E_i$ where $Y \sim N(0, \rho)$ and $E_i \sim N(0, 1-\rho)$. The E_i and Y are all independent. By standard ergodic theory (see Billingsley 1995), it follows that

$$\lim_{n\to\infty} \frac{\sum_{i=1}^{n} 1(Z_i \in \Gamma)}{n} \stackrel{a.s.}{=} \mathbf{E}[1(\mu + Y + E \in \Gamma) | \mathcal{I}]. \tag{58}$$

E has the same distribution as the E_i and it is measurable with respect to $\mathcal{I}$, the invariant sigma-field with respect to the shift operator. Thus,

$$\mathbf{E}[1(\mu + Y + E \in \Gamma) | \mathcal{I}] = \mathbf{E}_E[1(\mu + Y + E \in \Gamma)], \tag{59}$$

the expectation on the right hand side being taken over E. The c.d.f. of W_1 is equal to that of $\mathbf{E}_E[1(\mu + Y + E \in \Gamma)]$, and this can easily be calculated to equal the c.d.f. given in Example 6.2.

Acknowledgments

Thanks to Brad Efron, Rob Tibshirani, Larry Wasserman, and Ji Zhu for very helpful ideas and comments. Brad Efron contributed greatly to the ideas in Section 3 and Section 6. Also, thanks to David Siegmund for pointing out the Morton (1955) reference. This research was supported in part by a NSF Graduate Research Fellowship.

References

Abramovich F, 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.

Billingsley P. (1995) *Probability and Measure*, 3rd edition. New York: John Wiley and Sons.

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

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

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

Morton NE. (1955) Sequential tests for the detection of linkage. *American Journal of Human Genetics* **7:** 277-318.

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

Storey JD. (2001) A new approach to multiple hypothesis testing, in preparation.

Storey JD and Tibshirani RJ. (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.

Weller JI, Song JZ, Heyen DW, Lewin HA, and Ron M. (1998) A new approach to the problem of multiple comparisons in the genetic dissection of complex traits. *Genetics* **150:** 1699-1706.

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

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.

Zaykin DV, Young SS, and Westfall PH. (1998) Using the false discovery approach in the genetic dissection of complex traits: A response Weller et al. *Genetics* **154:** 1917-1918.