False Discovery Rates

John D. Storey
Princeton University, Princeton, USA
January 2010

Multiple Hypothesis Testing

In hypothesis testing, *statistical significance* is typically based on calculations involving p-values and Type I error rates. A p-value calculated from a single statistical hypothesis test can be used to determine whether there is statistically significant evidence against the null hypothesis. The upper threshold applied to the p-value in making this determination (often 5% in the scientific literature) determines the Type I error rate; i.e., the probability of making a Type I error when the null hypothesis is true. *Multiple hypothesis testing* is concerned with testing several statistical hypotheses simultaneously. Defining statistical significance is a more complex problem in this setting.

A longstanding definition of statistical significance for multiple hypothesis tests involves the probability of making one or more Type I errors among the family of hypothesis tests, called the *family-wise error rate*. However, there exist other well established formulations of statistical significance for multiple hypothesis tests. The Bayesian framework for classification naturally allows one to calculate the probability that each null hypothesis is true given the observed data (Efron et al. 2001, Storey 2003), and several frequentist definitions of multiple hypothesis testing significance are also well established (Shaffer 1995).

Soric (1989) proposed a framework for quantifying the statistical significance of multiple hypothesis tests based on the proportion of Type I errors among all hypothesis tests called statistically significant. He called statistically significant hypothesis tests *discoveries* and proposed that one be concerned about the rate of *false discoveries*[1] when testing multiple hypotheses. This false discovery rate is robust to the false positive paradox and is particularly useful in exploratory analyses, where one is more concerned with having mostly true findings among a set of statistically significant discoveries rather than guarding against one or more false positives. Benjamini & Hochberg (1995) provided the first implementation of false discovery rates with known operating characteristics. The idea of quantifying the rate of false discoveries is directly related to several pre-existing ideas, such as Bayesian misclassification rates and the positive predictive value (Storey 2003).

[1] A *false discovery*, *Type I error*, and *false positive* are all equivalent. Whereas the *false positive rate* and *Type I error rate* are equal, the *false discovery rate* is an entirely different quantity.

Applications

In recent years, there has been a substantial increase in the size of data sets collected in a number of scientific fields, including genomics, astrophysics, brain imaging, and spatial epidemiology. This has been due in part to an increase in computational abilities and the invention of various technologies, such as high-throughput biological devices. The analysis of high-dimensional data sets often involves performing simultaneous hypothesis tests on each of thousands or millions of measured variables. Classical multiple hypothesis testing methods utilizing the family-wise error rate were developed for performing just a few tests, where the goal is to guard against any single false positive occurring. However, in the high-dimensional setting, a more common goal is to identify as many true positive findings as possible, while incurring a relatively low number of false positives. The false discovery rate is designed to quantify this type of trade-off, making it particularly useful for performing many hypothesis tests on high-dimensional data sets.

Hypothesis testing in high-dimensional genomics data sets has been particularly influential in increasing the popularity of false discovery rates (Storey & Tibshirani 2003). For example, DNA microarrays measure the expression levels of thousands of genes from a single biological sample. It is often the case that microarrays are applied to samples collected from two or more biological conditions, such as from multiple treatments or over a time course. A common goal in these studies is to identify genes that are differentially expressed among the biological conditions, which involves performing a hypothesis tests on each gene. In addition to incurring false positives, failing to identify truly differentially expressed genes is a major concern, leading to the false discovery rate being in widespread use in this area.

Mathematical Definitions

Although multiple hypothesis testing with false discovery rates can be formulated in a very general sense (Storey 2007, Storey et al. 2007), it is useful to consider the simplified case where m hypothesis tests are performed with corresponding p-values $p_1, p_2, \ldots, p_m$. The typical procedure is to call null hypotheses statistically significant whenever their corresponding p-values are less than or equal to some threshold, $t \in (0, 1]$. This threshold can be fixed or data-dependent, and the procedure for determining the threshold involves quantifying a desired error rate.

Table 1 describes the various outcomes that occur when applying this approach to determining which of the m hypothesis tests are statistically significant. Specifically, V is the number of Type I errors (equivalently false positives or false discoveries) and R is the total number of significant null hypotheses (equivalently total discoveries). The *family-wise error rate* (FWER) is defined to be

$$\text{FWER} = \mathbf{Pr}(V \geq 1),$$

and the *false discovery rate* (FDR) is usually defined to be (Benjamini & Hochberg 1995):

$$\text{FDR} = \mathbf{E}\left[\frac{V}{R \vee 1}\right] = \mathbf{E}\left[\frac{V}{R}\middle| R > 0\right] \mathbf{Pr}(R > 0).$$

Table 1: *Possible outcomes from m hypothesis tests based on applying a significance threshold $t \in (0, 1]$ to their corresponding p-values.*

	Not Significant (p-value $> t$)	Significant (p-value $\leq t$)	Total
Null True	U	V	m_0
Alternative True	T	S	m_1
	W	R	m

The effect of "$R \vee 1$" in the denominator of the first expectation is to set $V/R = 0$ when $R = 0$. As demonstrated in Benjamini & Hochberg (1995), the FDR offers a less strict multiple testing criterion than the FWER, which is more appropriate for some applications.

Two other false discovery rate definitions have been proposed in the literature, where the main difference is in how the $R = 0$ event is handled. These quantities are called the *positive false discovery rate* (pFDR) and the *marginal false discovery rate* (mFDR), and they are defined as follows (Storey 2003, Storey 2007):

$$\text{pFDR} = \mathbf{E}\left[\frac{V}{R}\middle| R > 0\right],$$

$$\text{mFDR} = \frac{\mathbf{E}[V]}{\mathbf{E}[R]}.$$

Note that pFDR = mFDR = 1 whenever all null hypotheses are true, whereas FDR can always be made arbitrarily small because of the extra term $\mathbf{Pr}(R > 0)$. Some have pointed out that this extra term in the FDR definition may lead to misinterpreted results, and pFDR or mFDR offer more scientifically relevant values (Zaykin et al. 1998, Storey 2003); others have argued that FDR is preferable because it allows for the traditional strong control criterion to be met (Benjamini & Hochberg 1995). All three quantities can be utilized in practice, and they are all similar when the number of hypothesis tests is particularly large.

Control and Estimation

There are two approaches to utilizing false discovery rates in a conservative manner when determining multiple testing significance. One approach is to fix the acceptable FDR level beforehand, and find a data-dependent thresholding rule so that the expected FDR of this rule over repeated studies is less than or equal to the pre-chosen level. This property is called *FDR control* (Shaffer 1995, Benjamini & Hochberg 1995). Another approach is to fix the p-value threshold at a particular value and then form a point estimate of the FDR whose expectation is greater than or equal to the true FDR at that particular threshold (Storey 2002). The latter approach has been useful in that it places multiple testing in the more standard context of point estimation, whereas the derivation of

algorithms in the former approach may be less tractable. Indeed, it has been shown that the point estimation approach provides a comprehensive and unified framework (Storey et al. 2004).

For the first approach, Benjamini & Hochberg (1995) proved that the following algorithm for determining a data based p-value threshold controls the FDR at level α when the p-values corresponding to true null hypotheses are independent and identically distributed (i.i.d.) Uniform(0,1). Other p-value threshold determining algorithms for FDR control have been subsequently studied (e.g., Benjamini & Liu 1999). This algorithm was originally introduced by Simes (1986) to control the FWER when all p-values are independent and all null hypotheses are true, although it also provides control of the FDR for any configuration of true and false null hypotheses.

FDR Controlling Algorithm (Simes 1986; Benjamini & Hochberg 1995)

---

1. Let $p_{(1)} \leq \ldots \leq p_{(m)}$ be the ordered, observed p-values.
2. Calculate $\widehat{k} = \max\{1 \leq k \leq m : p_{(k)} \leq \alpha \cdot k/m\}$.
3. If $\widehat{k}$ exists, then reject null hypotheses corresponding to $p_{(1)} \leq \ldots \leq p_{(\widehat{k})}$. Otherwise, reject nothing.

To formulate the point estimation approach, let $\mathrm{FDR}(t)$ denote the FDR when calling null hypotheses significant whenever $p_i \leq t$, for $i = 1, 2, \ldots, m$. For $t \in (0, 1]$, we define the following stochastic processes based on the notation in Table 1:

$$\begin{aligned} V(t) &= \#\{\text{true null } p_i : p_i \leq t\}, \\ R(t) &= \#\{p_i : p_i \leq t\}. \end{aligned}$$

In terms of these, we have

$$\mathrm{FDR}(t) = \mathbf{E}\left[\frac{V(t)}{R(t) \vee 1}\right].$$

For fixed t, Storey (2002) provided a family of conservatively biased point estimates of $\mathrm{FDR}(t)$:

$$\widehat{\mathrm{FDR}}(t) = \frac{\widehat{m}_0(\lambda) \cdot t}{[R(t) \vee 1]}.$$

The term $\widehat{m}_0(\lambda)$ is an estimate of m_0, the number of true null hypotheses. This estimate depends on the tuning parameter λ, and it is defined as

$$\widehat{m}_0(\lambda) = \frac{m - R(\lambda)}{(1 - \lambda)}.$$

It can be shown that $\mathbf{E}[\widehat{m}_0(\lambda)] \geq m_0$ when the p-values corresponding to the true null hypotheses are Uniform(0,1) distributed (or stochastically greater). There is an inherent bias/variance trade-off in the choice of λ. In most cases, when λ gets smaller, the bias of $\widehat{m}_0(\lambda)$ gets larger, but the variance gets smaller. Therefore, λ can be chosen to try to balance this trade-off. Storey &

Tibshirani (2003) provide an intuitive motivation for the $\widehat{m}_0(\lambda)$ estimator, as well as a method for smoothing over the $\widehat{m}_0(\lambda)$ to obtain a tuning parameter free $\widehat{m}_0$ estimator. Sometimes instead of m_0, the quantity $\pi_0 = m_0/m$ is estimated, where simply $\widehat{\pi}_0(\lambda) = \widehat{m}_0(\lambda)/m$.

To motivate the overall estimator $\widehat{\text{FDR}}(t) = \widehat{m}_0(\lambda) \cdot t/[R(t) \vee 1]$, it may be noted that $\widehat{m}_0(\lambda) \cdot t \approx V(t)$ and $[R(t) \vee 1] \approx R(t)$. It has been shown under a variety of assumptions, including those of Benjamini & Hochberg (1995), that the desired property $\mathbf{E}\left[\widehat{\text{FDR}}(t)\right] \geq \text{FDR}(t)$ holds.

Storey et al. (2004) have shown that the two major approaches to false discovery rates can be unified through the estimator $\widehat{\text{FDR}}(t)$. Essentially, the original FDR controlling algorithm can be obtained by setting $\widehat{m}_0 = m$ and utilizing the p-value threshold $t^*_\alpha = \max\left\{t : \widehat{\text{FDR}}(t) \leq \alpha\right\}$. By allowing for the different estimators $\widehat{m}_0(\lambda)$, a family of FDR controlling procedures can be derived in this manner. In the asymptotic setting where the number of tests m is large, it has also been shown that the two approaches are essentially equivalent.

Q-values

In single hypothesis testing, it is common to report the p-value as a measure of significance. The "q-value" is the FDR based measure of significance that can be calculated simultaneously for multiple hypothesis tests. Initially it seems that the q-value should capture the FDR incurred when the significance threshold is set at the p-value itself, $\text{FDR}(p_i)$. However, unlike Type I error rates, the FDR is not necessarily strictly increasing with an increasing significance threshold. To accommodate this property, the q-value is defined to be the minimum FDR (or pFDR) at which the test is called significant (Storey 2002, Storey 2003):

$$\text{q-value}(p_i) = \min_{t \geq p_i} \text{FDR}(t) \quad \text{or} \quad \text{q-value}(p_i) = \min_{t \geq p_i} \text{pFDR}(t).$$

To estimate this in practice, a simple plug-in estimate is formed, for example:

$$\widehat{\text{q}}\text{-value}(p_i) = \min_{t \geq p_i} \widehat{\text{FDR}}(t).$$

Various theoretical properties have been shown for these estimates under certain conditions, notably that the estimated q-values of the entire set of tests are simultaneously conservative as the number of hypothesis tests grows large (Storey et al. 2004).

Bayesian Derivation

The pFDR has been shown to be exactly equal to a Bayesian derived quantity measuring the probability that a significant test is a true null hypothesis. Suppose that (a) $H_i = 0$ or 1 according to whether the ith null hypothesis is true or not, (b) $H_i \overset{i.i.d.}{\sim} \text{Bernoulli}(1-\pi_0)$ so that $\mathbf{Pr}(H_i = 0) = \pi_0$ and $\mathbf{Pr}(H_i = 1) = 1 - \pi_0$, and (c) $P_i | H_i \overset{i.i.d.}{\sim} (1 - H_i) \cdot G_0 + H_i \cdot G_1$, where G_0 is the null distribution

and G_1 is the alternative distribution. Storey (2001, 2003) showed that in this scenario

$$\begin{aligned} \text{pFDR}(t) &= \mathbf{E}\left[\frac{V(t)}{R(t)}\middle| R(t) > 0\right] \\ &= \mathbf{Pr}(H_i = 0 | P_i \leq t), \end{aligned}$$

where $\mathbf{Pr}(H_i = 0|P_i \leq t)$ is the same for each i because of the i.i.d. assumptions. Under these modeling assumptions, it follows that q-value$(p_i) = \min_{t \geq p_i} \mathbf{Pr}(H_i = 0|P_i \leq t)$, which is a Bayesian analogue of the p-value – or rather a "Bayesian posterior Type I error rate." Related concepts were suggested as early as Morton (1955). In this scenario, it also follows that $\text{pFDR}(t) = \int \mathbf{Pr}(H_i = 0|P_i = p_i) dG(p_i|p_i \leq t)$, where $G = \pi_0 G_0 + (1 - \pi_0) G_1$. This connects the pFDR to the posterior error probability $\mathbf{Pr}(H_i = 0|P_i = p_i)$, making this latter quantity sometimes interpreted as a *local false discovery rate* (Efron et al. 2001, Storey 2001).

Dependence

Most of the existing procedures for utilizing false discovery rates in practice involve assumptions about the p-values being independent or weakly dependent. An area of current research is aimed at performing multiple hypothesis tests when there is dependence among the hypothesis tests, specifically at the level of the data collected for each test or the p-values calculated for each test. Recent proposals suggest modifying FDR controlling algorithms or extending their theoretical characterizations (Benjamini & Yekutieli 2001), modifying the null distribution utilized in calculating p-values (Devlin & Roeder 1999, Efron 2004), or accounting for dependence at the level of the originally observed data in the model fitting (Leek & Storey 2007, Leek & Storey 2008).

References

Benjamini, Y. & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing, *Journal of the Royal Statistical Society, Series B* **85**: 289–300.

Benjamini, Y. & Liu, W. (1999). A step-down multiple hypothesis procedure that controls the false discovery rate under independence, *J. Stat. Plan. and Inference* **82**: 163–170.

Benjamini, Y. & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency, *Annals of Statistics* **29**: 1165–1188.

Devlin, B. & Roeder, K. (1999). Genomic control for association studies, *Biometrics* **55**: 997–1004.

Efron, B. (2004). Large-scale simultaneous hypothesis testing: The choice of a null hypothesis, *Journal of the American Statistical Association* **99**: 96–104.

Efron, B., Tibshirani, R., Storey, J. D. & Tusher, V. (2001). Empirical Bayes analysis of a microarray experiment, *Journal of the American Statistical Association* **96**: 1151–1160.

Leek, J. T. & Storey, J. D. (2007). Capturing heterogeneity in gene expression studies by surrogate variable analysis, *PLoS Genetics* **3**: e161.

Leek, J. T. & Storey, J. D. (2008). A general framework for multiple testing dependence., *Proceedings of the National Academy of Sciences* **105**: 18718–18723.

Morton, N. E. (1955). Sequential tests for the detection of linkage, *American Journal of Human Genetics* **7**: 277–318.

Shaffer, J. (1995). Multiple hypothesis testing, *Annual Rev. Psych.* **46**: 561–584.

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

Soric, B. (1989). Statistical discoveries and effect-size estimation, *Journal of the American Statistical Association* **84**: 608–610.

Storey, J. D. (2001). The positive false discovery rate: A Bayesian interpretation and the q-value. Technical Report 2001-12, Department of Statistics, Stanford University.

Storey, J. D. (2002). A direct approach to false discovery rates, *Journal of the Royal Statistical Society, Series B* **64**: 479–498.

Storey, J. D. (2003). The positive false discovery rate: A Bayesian interpretation and the q-value, *Annals of Statistics* **31**: 2013–2035.

Storey, J. D. (2007). The optimal discovery procedure: A new approach to simultaneous significance testing, *Journal of the Royal Statistical Society, Series B* **69**: 347–368.

Storey, J. D., Dai, J. Y. & Leek, J. T. (2007). The optimal discovery procedure for large-scale significance testing, with applications to comparative microarray experiments, *Biostatistics* **8**: 414–432.

Storey, J. D., Taylor, J. E. & Siegmund, D. (2004). Strong control, conservative point estimation, and simultaneous conservative consistency of false discovery rates: A unified approach, *Journal of the Royal Statistical Society, Series B* **66**: 187–205.

Storey, J. D. & Tibshirani, R. (2003). Statistical significance for genome-wide studies, *Proceedings of the National Academy of Sciences* **100**: 9440–9445.

Zaykin, D. V., Young, S. S. & Westfall, P. H. (1998). Using the false discovery approach in the genetic dissection of complex traits: A response to weller et al., *Genetics* **150**: 1917–1918.