9-6-2005

The Optimal Discovery Procedure: A New Approach to Simultaneous Significance Testing

John D. Storey
University of Washington, jstorey@u.washington.edu

The Optimal Discovery Procedure: A New Approach to Simultaneous Significance Testing

John D. Storey
Department of Biostatistics
University of Washington
jstorey@u.washington.edu

Abstract: Significance testing is one of the main objectives of statistics. The Neyman-Pearson lemma provides a simple rule for optimally testing a single hypothesis when the null and alternative distributions are known. This result has played a major role in the development of significance testing strategies that are used in practice. Most of the work extending single testing strategies to multiple tests has focused on formulating and estimating new types of significance measures, such as the false discovery rate. These methods tend to be based on p-values that are calculated from each test individually, ignoring information from the other tests. As shrinkage estimation borrows strength across point estimates to improve their overall performance, I show here that borrowing strength across multiple significance tests can improve their performance as well. The "optimal discovery procedure" (ODP) is introduced, which shows how to maximize the number of expected true positives for each fixed number of expected false positives. The optimality achieved by this procedure is shown to be closely related to optimality in terms of the false discovery rate. The ODP motivates a new approach to testing multiple hypotheses, especially when the tests are related. As a simple example, a new simultaneous procedure for testing several Normal means is defined; this is surprisingly demonstrated to outperform the optimal single test procedure, showing that an optimal method for single tests may no longer be optimal in the multiple test setting. Connections to other concepts in statistics are discussed, including Stein's paradox, shrinkage estimation, and Bayesian classification theory.

1 Introduction

In 1933, Jerzy Neyman and Egon Pearson derived the optimal procedure for performing a single significance test when the null and alternative distributions are known (Neyman & Pearson 1933). Given observed data, the optimal testing procedure is based on the likelihood ratio

$$\frac{\text{likelihood of data under alternative distribution}}{\text{likelihood of data under null distribution}}.$$

The null hypothesis is then rejected if the likelihood ratio exceeds some pre-chosen cut-off. This Neyman-Pearson (NP) procedure is optimal because it is "most powerful," meaning that for each fixed Type I error rate, there does not exist another rule that exceeds this one in power. The optimality follows intuitively from the fact that the strength of the alternative versus the null is assessed by comparing their exact likelihoods.

Here I consider the situation where multiple significance tests are simultaneously performed, a process typically involving two major steps: (1) rank the tests from most to least significant and (2) choose an appropriate significance cut-off somewhere along this ranking. More formally, in the first step one must derive a rule for rejecting null hypotheses, an analogous problem to the one that Neyman and Pearson considered. Such a rule is usually defined in terms of a statistic that has been calculated for each test, where a single significance cut-off is applied to the statistics. This first step can be thought of very simply as indicating *which set* of null hypotheses should be rejected at each possible significance level. Given this general rule, the second step is to estimate the significance level associated with each cut-off so that a specific threshold can be set to attain an acceptable error rate.

The field of "multiple hypothesis testing" has been focused on the second step: formulating and estimating multiple testing error rates, such as the family-wise error rate or false discovery rate, once the general significance rule has already been defined (Shaffer 1995). Multiple testing methods are typically defined in terms of p-values that have individually been obtained from each significance test, using information from only that significance test. In such a setting, the first step is simple: call all tests significant with p-values less than or equal to some cut-off, where each test has been considered separately in order to obtain these p-values. The goal is then to estimate the appropriate cut-off for obtaining a particular error rate. However, notice that when taking this approach, no information in the data is *across tests* when assessing how relatively significant each test is.

The problem of choosing from among some pre-defined set of significance cut-offs in order to "control" an error rate is not considered here. Rather, I focus on defining the set of significance cut-offs themselves in order to optimally perform the multiple tests at all significance levels, analogous

to Neyman and Pearson's goals for a single significance test. In other words, I show *which* null hypotheses to reject at each possible significance level in order to optimize the overall testing procedure. An extension of the Type I error rate and power is considered that is closely related to a variety of multiple testing error rates. The specific goal is to maximize the expected number of true positives for each fixed expected number of false positives. I derive and investigate the significance rule, called the "optimal discovery procedure" (ODP) lemma, that achieves this optimization. This procedure involves the formation of a statistic for each test that uses the relevant information from every other test, similar to shrinkage estimators now commonly used in simultaneous point estimation.

The ODP assumes that the true probability densities corresponding to each significance test are known, which is analogous to the assumption made for the original Neyman-Pearson (NP) lemma. In practice, these probability densities will typically be unknown. Therefore, an approach is briefly illustrated to show that the ODP may be applied to derive practical multiple testing procedures. This application of the ODP is fully developed and applied in a subsequent paper (Storey et al. 2005). The ODP is shown to have connections to Bayesian classification theory, although it remains a truly frequentist procedure. In fact, it is shown that Bayesian approaches to simultaneous testing may be more closely related to the NP approach than the ODP. I also show that the optimality achieved by the ODP procedure is equivalent to an optimality under the false discovery rate (FDR) error measure.

The ODP shows that the overall performance of multiple significance tests can be improved by "borrowing strength" across these tests. This is similar to Stein's paradox shown in 1956 for point estimation (Stein 1956, Stein 1981). Stein showed that the estimation of several Normal means can be improved by using information from all data simultaneously. It was surprisingly shown that the point estimate that is optimal for a single parameter is no longer optimal when estimating multiple parameters. Here I empirically show a similar result where several Normals means are tested for equality to zero. By using a simple procedure well motivated by the ODP lemma, I show that the uniformly most powerful unbiased (UMP unbiased) test of a Normal mean is no longer uniformly most powerful in the multivariate setting. That is, a procedure considered to be optimal for a single test is no longer optimal when applied to several tests simultaneously.

2 Motivating Example: Testing Several Normal Means

The ideas introduced in this paper can be motivated through the following example where multiple Normal means are tested. This example is employed throughout the paper.

2.1 Testing a single Normal mean

First consider a single significance test performed on the mean of a Normal distribution, based on a single observation z from the $N(\mu, 1)$ distribution. Suppose that the the null hypothesis is $\mu = 0$ and the alternative is $\mu = 2$. Any significance testing procedure can be described in terms of a significance thresholding function $\mathcal{S}(\text{data})$, defined so that the null hypothesis is rejected (i.e., the test is significant) if $\mathcal{S}(\text{data}) \geq \lambda$ (Lehmann 1986). The value λ is determined to satisfy a user-chosen Type I error rate. In this example, the observation z is used to decide between the two hypotheses. The NP lemma implies that the most powerful testing procedure is based on the significance thresholding function defined to be the ratio of the likelihood of z under the alternative to the likelihood under the null:

$$\mathcal{S}(z) = \frac{\frac{1}{\sqrt{2\pi}} \exp \frac{-(z-2)^2}{2}}{\frac{1}{\sqrt{2\pi}} \exp \frac{-z^2}{2}}.$$

The optimal thresholding rule $\mathcal{S}(z) \geq \lambda$ can be shown to be equivalent to $z \geq c$ for some c yielding an equivalent Type I error rate. Note that the same rule would emerge for any alternative value of μ greater than zero. Therefore, as long as the value of μ under the alternative is greater than zero, then $z \geq c$ is always the most powerful rule. When the value of μ under the alternative is less than zero, the most powerful rule is equivalent to rejecting the null hypothesis when $z \leq c$.

In practice, the alternative value of μ is usually unknown, making it necessary to perform a two-sided significance test. In this case, the null hypothesis is $\mu = 0$ and the alternative is $\mu \neq 0$. The Neyman-Pearson approach has been extended in several ways to deal with this uncertainty. One strategy is to effectively estimate the NP rule. Let $\widehat{\mu}$ be an estimate of the mean μ under the alternative hypothesis, say $\widehat{\mu} = z$. An estimated version of the NP thresholding function is

$$\widehat{\mathcal{S}}(z) = \frac{\frac{1}{\sqrt{2\pi}} \exp \frac{-(z-\widehat{\mu})^2}{2}}{\frac{1}{\sqrt{2\pi}} \exp \frac{-z^2}{2}}.$$

It can be shown that rejecting the null hypothesis when $\widehat{\mathcal{S}}(z) \geq \lambda$ is equivalent to employing the threshold $|z| \geq c$ for some equivalent cut-off c. As the number of observations for the test grows large, this procedure is the asymptotically optimal "generalized likelihood ratio test" (Lehmann 1986). Another approach, also building on the NP lemma, is to find the most powerful rule among all of those that are unbiased (i.e., where the power is greater than or equal to the Type I error for every value of the alternative hypothesis). As it turns out, rejecting the null hypothesis when $|z| \geq c$ is also the UMP unbiased procedure for the test of $\mu = 0$ versus $\mu \neq 0$ (Lehmann 1986).

2.2 Testing several Normal means

Now consider the case where multiple significance tests are performed. Suppose that we observe $z_i \sim N(\mu_i, 1)$ for tests $i = 1, \ldots, 8$, where significance test i is $\mu_i = 0$ versus $\mu_i \neq 0$. The conventional approach to testing multiple hypotheses is to apply the best single test rule to each one individually. In our example, this involves fixing a cut-off c and rejecting all null hypotheses with $|z_i| \geq c$. Since each test has the same null distribution, this is equivalent to calculating a p-value for each test and then forming the equivalent p-value threshold. Therefore, even though the true NP rule allows the significance rule to differ between tests (i.e., $z_i \leq c$ or $z_i \geq c$, depending on the sign of the alternative parameter value), the usual, practical version leads to a common significance threshold among all tests. It can be verified that this is often the case in standard situations, for example, when performing multiple chi-square tests or two-sided t-tests. This motivates one to determine the optimal rule among all "simultaneous thresholding procedures", i.e., those that apply one significance rule to all tests.

3 Optimal Discovery Procedure: Theory

I now introduce the optimal procedure among all simultaneous thresholding rules for performing multiple significance tests, called the "optimal discovery procedure" (ODP).

3.1 Optimality goal

As a multivariate extension of maximizing power for each fixed Type I error rate, *the optimality goal I propose is to maximize the expected number of true positives (ETP) for each fixed expected number of false positives (EFP).* Recall that each test has a Type I error rate level and a power level. For any true null hypothesis, the Type I error rate is the expected number of false positives resulting from that test. Therefore, the sum of Type I error rates across all true null hypotheses is the EFP. Similarly, the sum of powers across all true alternative hypotheses is the ETP. Even though there are many ways in which one can combine these values in order to assess the performance of multiple significance tests, this particular one is focused on the overall "discovery rate" (Soric 1989). Finally, note that the EFP-to-ETP trade-off proposed here is easily extended to the case where each test is given a specific relative weight, making it unnecessary to treat every positive equally (see the remark following the statement of Lemma 1).

The proposed optimality criterion is also directly related to optimality in terms of FDRs and misclassification rates. For FDRs, the key observation is that

$$\text{FDR} \approx \frac{\text{EFP}}{\text{EFP} + \text{ETP}}, \tag{1}$$

where the approximate equality is even sometimes an exact one. An exact equality exists for large numbers of tests with certain convergence properties (Storey et al. 2004), under Bayesian mixture model assumptions (Storey 2003), and under alternative definitions of the FDR (Benjamini & Hochberg 1995, Storey 2003). The important point is that the FDR may be interpreted and characterized in terms of the EFP and ETP.

It is shown in Section 6 that if the ETP is maximized for each fixed EFP level, then the proportion of "missed discoveries" (Genovese & Wasserman 2002, Taylor et al. 2005) is minimized for each fixed FDR level; that is, achieving optimality in terms of the EFP and ETP is equivalent to achieving optimality in terms of the FDR. (This exact statement is sometimes an approximate one, depending on the relationship between the two quantities in equation (1).) Optimality in terms of misclassification error is also achieved under the goal of maximizing ETP for each fixed EFP level. Therefore, even though I derive the theory below in terms of the EFP and ETP, it can also be applied in terms of these other error measures. Furthermore, it may be argued that the EFP and ETP are the more fundamental units of a number of error measures, making them an appropriate choice for defining optimality.

As stated above, a significance testing procedure can be described in terms of a significance thresholding function $\mathcal{S}(\text{data})$, defined so that the null hypothesis is rejected if $\mathcal{S}(\text{data}) \geq \lambda$. When multiple tests are performed, it may be the case that a separate $\mathcal{S}$ is defined for each test. In such a case, a threshold would be applied to $\mathcal{S}_i(\text{data}_i)$, where $\mathcal{S}_i$ is the test-specific thresholding function and data_i are the data for test i. A single, simultaneous thresholding procedure is one where a single thresholding function is applied to all tests. Due to the fact that most multiple testing procedures can be written in terms of a single, simultaneous thresholding procedure, the ODP is defined to be the one that maximizes the ETP for each fixed EFP level among all simultaneous thresholding procedures. Section A.1 shows that most multiple testing procedures are in fact equivalent to applying a single, simultaneous thresholding rule, strongly motivating the optimality criterion defined here.

3.2 ODP for testing several Normal means

I first present the ODP for the multiple Normal means example introduced above. First, assume that we know the *true* values of the means $\mu_1, \mu_2, \ldots, \mu_8$. This implies that we know whether each null hypothesis is true or false, and if it is false, we know the alternative distribution. Table 1 provides this information for the eight significance tests. According to Lemma 1 below, the ODP is based on the following significance thresholding function:

$$\mathcal{S}_{\text{ODP}}(z) = \frac{\text{sum of the true alternative densities evaluated at } z}{\text{sum of true null densities evaluated at } z}. \tag{2}$$

Table 1: An example of eight significance tests performed on the mean of a Normal distribution, each based on a single observation z_i from the $N(\mu_i, 1)$ distribution, $i = 1, 2, \ldots, 8$. For each test, the the null hypothesis is $\mu_i = 0$. The first row of the table gives the value of μ_i under the alternative if it were known. The second row is the *true* value of μ_i. The third row is the observation for each test, z_i. The fourth row gives the ranking of the tests in terms of their signficance according to the ODP. The fifth row gives the ranking based on the estimated ODP, which tests the alternative hypothesis $\mu_i \neq 0$. The sixth row is the significance ranking based on the univariate UMP unbiased test against the alternative $\mu_i \neq 0$, which uses the significance thresholding rule $|z_i| \geq c$.

Significance test $i =$	1	2	3	4	5	6	7	8
Alternative value of μ_i	-3	-2	-2	-1	1	2	2	3
True value of μ_i	0	-2	0	0	1	2	0	3
Observed datum, z_i	1.0	-2.3	-0.02	-0.4	0.5	2.2	-0.1	3.4
ODP significance rank	4	3	6	8	5	2	7	1
Estimated ODP signif. rank	4	3	6	8	5	2	7	1
UMP unbiased signif. rank	4	2	8	6	5	3	7	1

This is similar to the NP likelihood ratio, except the likelihood of each test's data is considered under the true probability density function of every test. Evidence is added across the true alternatives and the true nulls in forming the ratio. This makes sense in that we are trying to optimize the procedure over all tests.

Letting $\phi(z;\mu) = \frac{1}{\sqrt{2\pi}} \exp\{-(z-\mu)^2/2\}$ be the density of a $N(\mu, 1)$ random variable, the significance thresholding function can be written more formally as

$$\mathcal{S}_{\text{ODP}}(z) = \frac{\phi(z;-2) + \phi(z;1) + \phi(z;2) + \phi(z;3)}{\phi(z;0) + \phi(z;0) + \phi(z;0) + \phi(z;0)}. \tag{3}$$

For a fixed λ chosen to attain an acceptable EFP level, null hypothesis i is rejected if $\mathcal{S}_{\text{ODP}}(z_i) \geq \lambda$. The observed data for test i has been evaluated at the true densities among *all* tests. These are summed and compared to the null density in order to decide whether to call test i significant or not.

The intuition behind this procedure is that the significance of each observed z_i is made relatively higher as evidence builds up that there are multiple true positives likely to have similar values. Supposing that z_i corresponds to a true alternative hypothesis, it should be the case that z_i is close to μ_i with high probability, making the contribution from its density $\frac{1}{\sqrt{2\pi}} \exp \frac{-(z_i-\mu_i)^2}{2}$ to $\mathcal{S}_{\text{ODP}}(z_i)$ substantial. However, if there are other true alternatives with $\mu_j \approx \mu_i$, then the likelihood of z_i

under $\frac{1}{\sqrt{2\pi}} \exp \frac{-(z_i-\mu_j)^2}{2}$ will also make a substantial contribution to $\mathcal{S}_{\text{ODP}}(\mu_j)$. Since the goal is to maximize the ETP for each fixed EFP, it makes sense to increase the relative significance of a particular test if there are other hypotheses with similar signal that also are well distinguished from the true null hypotheses. If one particular true alternative has mean μ_i very different than the others, then $\mathcal{S}_{\text{ODP}}(z_i)$ will behave like its NP statistic because the contribution from the other densities will be negligible.

Figure 1a shows a plot of $\mathcal{S}_{\text{ODP}}(z)$ over a range of z values for this particular example. It can be seen that $\mathcal{S}_{\text{ODP}}(z)$ captures the asymmetry in the signal among the true alternative hypotheses. The true alternative mean values are -2, 1, 2 and 3, so the significance is increased among the positive values of z. Table 1 gives an example of realized values of $z_1, z_2, \ldots, z_8$. From Table 1 it can be seen that the statistic for test 6 is greater than that for test 2, i.e., $\mathcal{S}_{\text{ODP}}(2.2) > \mathcal{S}_{\text{ODP}}(-2.3)$. The UMP unbiased procedure uses a symmetric significance rule, which is also shown in Figure 1a. Under this rule, test 2 with $|z_2| = 2.3$ would be considered more significant than test 6 with $|z_6| = 2.2$. The ODP significance rule ranks test 6 higher than test 2 because the true alternative means 1, 2, and 3 all contribute substantially to $\mathcal{S}_{\text{ODP}}(2.2)$.

It seems paradoxical to define the ODP under the assumption that the truth about each hypothesis is known. However, it is shown in the next section and in a paper applying this theory (Storey et al. 2005) that our theoretical result allows for a straightforward approach to estimating the ODP, requiring no estimation beyond what is required for estimating the NP rule for a single significance test.

3.3 ODP in a general setting

The ODP can be formulated and its optimality proven in a general setting. The tests are based on observed data sets $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m$, where $\mathbf{x}_i$ corresponds to significance test i. In general it should be assumed that the data sets are all random vectors defined on a common probability space. For simplicity one can think of the data sets as being composed of n observations each, $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{in})^T$. The following lemma defines the ODP in a general setting and proves its optimality.

Lemma 1: Optimal Discovery Procedure. *Suppose that m simple significance tests are performed on observed data sets $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m$, where significance test i has null density f_i and alternative density g_i, for $i = 1, \ldots, m$. Without loss of generality suppose that the null hypothesis is true for tests $i = 1, 2, \ldots, m_0$, and the alternative is true for $i = m_0 + 1, \ldots, m$. The following*

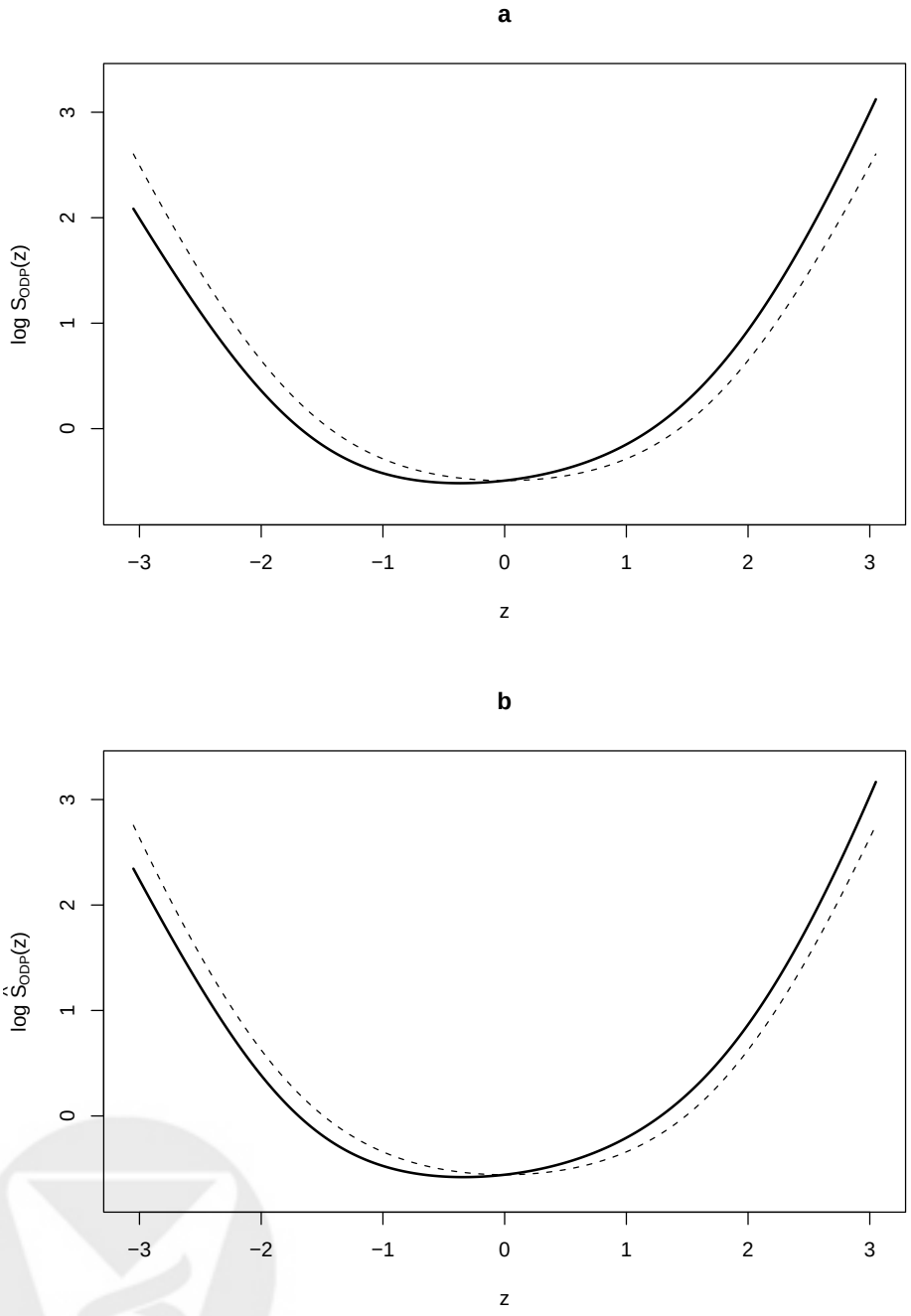

Figure 1: Plots of the true and estimated ODP significance thresholding functions for the multiple Normal means summarized in Table 1. **(a)** A plot of the significance thresholding function $\mathcal{S}_{\text{ODP}}(z)$ versus observed statistic z for true ODP (black) and a symmetric thresholding function (dash), which is equivalent to the optimal procedure when performing a single test. **(b)** Analogous plot to (a), except the estimated ODP significance thresholding function as defined in equation (7) is shown.

significance thresholding function defines the ODP:

$$\mathcal{S}_{\text{ODP}}(\mathbf{x}) = \frac{g_{m_0+1}(\mathbf{x}) + g_{m_0+2}(\mathbf{x}) + \cdots + g_m(\mathbf{x})}{f_1(\mathbf{x}) + f_2(\mathbf{x}) + \cdots + f_{m_0}(\mathbf{x})}. \tag{4}$$

Null hypothesis i is rejected if and only if $\mathcal{S}(\mathbf{x}_i) \geq \lambda$ for some $0 \leq \lambda < \infty$. For each fixed λ, this procedure yields the maximum number of expected true positives (ETP) among all simultaneous thresholding procedures that have an equal or less number of expected false positives (EFP).

The proof of the lemma is in Section A.2. Note that no restrictions are placed on the probabilistic dependence between the tests; that is, the lemma holds under arbitrary dependence. Below, we extend the ODP to the case where the status of each significance test is random, producing a rule that uses both the null and alternative densities of every test.

Remark. It should be noted that it is possible to give each test a relative weight. For example, if it were known that a set of tests were related, then these could be weighted so that they essentially count as a single false positive or a single true positive. As another example, if prior information on certain tests indicate that these are more important, then they can be given higher weight. In general, if test i is given relative weight w_i, then the ODP can be generalized to maximize this weighted version of the ETP in terms of the weighted version of the EFP. The only difference in the formula above is that each f_i or g_i is multiplied by w_i. A positive from test i then contributes w_i to the EFP or ETP. The proof of this easily follows from the proof of Lemma 1 in Section A.2.

4 Optimal Discovery Procedure: Estimation

The ODP as presented above requires one to know the true distribution corresponding to each significance test, but this will not be known in practice. However, these may be estimated because the data observed for each test do in fact come from their true distribution, whether it be a null or alternative distribution. Therefore, one does not necessarily need to know the status of each test in order to effectively estimate the ODP. The following is one strategy for doing so in the multiple Normal means example; this has also been generalized for more complicated scenarios (Storey et al. 2005).

4.1 Estimated ODP for testing several Normal means

In the multiple Normal means example summarized in Table 1, the ODP thresholding function is shown in equations (2) and (3). Letting $\mu_1, \mu_2, \ldots, \mu_8$ denote the true means, the thresholding

function is

$$\mathcal{S}_{\text{ODP}}(z) = \frac{\phi(z;\mu_2) + \phi(z;\mu_5) + \phi(z;\mu_6) + \phi(z;\mu_8)}{\phi(z;\mu_1) + \phi(z;\mu_3) + \phi(z;\mu_4) + \phi(z;\mu_7)}. \tag{5}$$

In practice, every μ_i would be unknown, and the fact that there are four true null hypotheses in the denominator would also be unknown. *However, as it turns out, in order to estimate the ODP in this example, all we need to do is estimate the true μ_i for each test; it is not necessary to distinguish the true and false null hypotheses.* If it were necessary to distinguish the true and false null hypotheses in order to estimate the ODP, then it would be of no practical use.

Some rearranging of the function in equation (5) yields a form that is estimable without requiring any explicit separation of true and false null hypotheses. Note that the threshold $\mathcal{S}_{\text{ODP}}(z) \geq \lambda$ is equivalent to $[\mathcal{S}_{\text{ODP}}(z) + 1] \geq [\lambda + 1]$. It can easily be calculated that

$$\mathcal{S}_{\text{ODP}}(z) + 1 = \frac{\sum_{i=1}^{8} \phi(z;\mu_i)}{4 \times \phi(z;0)},$$

where the denominator follows from the fact that all true null hypothesis follow the $N(0,1)$ distribution. In practice, one may estimate $\sum_{i=1}^{8} \phi(z;\mu_i)$ by estimating each individual μ_i and then forming a plug-in estimate of the total sum. The "4" in the denominator (also unknown in practice) is not actually necessary. Since all test statistics have the same denominator, rescaling them all by 4 does not affect their significance rankings; in mathematical terms, $4[\mathcal{S}_{\text{ODP}}(z) + 1]$ is simply a monotone transformation of $\mathcal{S}(z)$, where $4[\mathcal{S}_{\text{ODP}}(z)+1]$ results in the function $\sum_{i=1}^{8} \phi(z;\mu_i)/\phi(z;0)$.

Therefore, the ODP can equivalently be performed with the thresholding function,

$$\mathcal{S}^*_{\text{ODP}}(z) = \frac{\sum_{i=1}^{8} \phi(z;\mu_i)}{\phi(z;0)}, \tag{6}$$

Since $\mathcal{S}^*_{\text{ODP}} = 4\mathcal{S}_{\text{ODP}} + 4$, the thresholding function in equation (6) is strictly monotone with respect to the thresholding function in equation (5), implying that they yield equivalent testing procedures. The main point of this manipulation is to show that the ODP can be written so that only the μ_i need to be estimated, where the true and false null hypotheses do not need to be directly distinguished. Even though the derivation in this example uses the fact that $\mu_1 = \mu_3 = \mu_4 = \mu_7 = 0$, the thresholding function in equation (6) will result for any combination of true and false null hypotheses.

Each true mean can be estimated by the data observed for its respective test. For example, we can set $\widehat{\mu}_j = z_j$ and substitute these into $\mathcal{S}^*_{\text{ODP}}(z)$, producing an estimated version of the ODP rule:

$$\widehat{\mathcal{S}}^*_{\text{ODP}}(z) : \frac{\sum_{i=1}^{m} \phi(z;\widehat{\mu}_i)}{\phi(z;0)}. \tag{7}$$

Therefore, for a fixed λ, test j is called significant if $\sum_{i=1}^{m} \phi(z_j; \widehat{\mu}_i)/\phi(z_j; 0) \geq \lambda$.

Table 1 gives a simulated example of realized values of $z_1, z_2, \ldots, z_8$. Substituting these values into the estimate, we get

$$\widehat{\mathcal{S}}^*_{\text{ODP}}(z) = \frac{\phi(z; 1.0) + \phi(z; -2.3) + \phi(z; -.02) + \phi(z; -.4) + \phi(z; .5) + \phi(z; 2.2) + \phi(z; -.1) + \phi(z; 3.4)}{\phi(z; 0)}.$$

Figure 1b shows a plot of $\widehat{\mathcal{S}}^*_{\text{ODP}}(z)$ versus z as well as a plot of the symmetric thresholding function, which is equivalent to the single test UMP unbiased rule. It can be seen that the asymmetry in the distribution of the μ_i's is captured in this estimated rule, and that the rule is similar to the true ODP shown in Figure 1a.

Simulated data sets such as the one above can be generated to numerically calculate the ETP level for each fixed EFP level. The procedure can then be compared to the UMP unbiased test that employs the symmetric significance rule $|z_i| \geq c$. Examples are shown below where the estimated ODP is superior (in terms of EFP versus ETP) to the UMP unbiased test for single significance tests, implying that optimal single test procedures may no longer be optimal in the multiple test setting.

4.2 Estimated ODP in a general setting

Two simplifying properties of the above ODP estimate are that (i) every test has the same null distribution and (ii) there are no nuisance parameters. It is possible to estimate the ODP in much more general scenarios. In a paper applying the theory presented here, we develop some general strategies for estimating the ODP, and we apply it the problem of identifying differentially expressed genes in DNA microarray experiments (Storey et al. 2005). This involves formulating an estimate when every test may have a different probability distribution, including the possibility that every null distribution is different. The approach is related to that above, although a number of additional complications must be properly handled. Nevertheless, it is shown there that the theory developed here can be applied to substantially more complicated scenarios.

5 Estimated ODP Versus UMP Unbiased Test

The performance of the estimated ODP is now compared to the UMP unbiased procedure in the multiple Normal means example. Recall that the UMP unbiased procedure rejects null hypotheses with $|z_i| \geq c$ for some fixed c. The thresholding rule for the estimated ODP is given in equation (7), where each null hypothesis is rejected if $\widehat{\mathcal{S}}^*_{\text{ODP}}(z) \geq \lambda$ for some fixed λ. Figure 2 shows a comparison among the true ODP, the estimated ODP, and the UMP unbiased procedure in terms of the ETP for each fixed EFP level. These comparisons were calculated via simulation, where the number of

iterations was large enough that the Monte Carlo error is negligible.

The six plots in Figure 2 show comparisons over a variety of configurations of true alternative means and numbers of tests performed. It can be seen that the true ODP is always the best performer, as the theory implies it should be. The larger the number of tests, and the larger the proportion of true alternatives to true nulls, the closer the estimated ODP is to the true ODP in terms of performance. The estimated ODP rule outperforms the UMP unbiased test in all cases where the alternative means are not arranged in a perfectly symmetric fashion around zero. What is meant by perfect symmetry is that if there is a true alternative mean of value μ_i, then there is another true alternative mean of value $-\mu_i$. When there is perfect symmetry, it can be shown that the true ODP and the UMP unbiased procedure are equivalent; in this case, the estimated ODP is slightly inferior to the other two because it is a noisy version of the perfectly symmetric rule.

What is notable about these results is that the UMP unbiased test is no longer optimal in the multiple testing setting. In other words, *an optimal single testing procedure may no longer be optimal when performing multiple tests.* The UMP unbiased test is "uniformly most powerful" for a single significance test of $\mu = 0$ versus $\mu \neq 0$. This means that for any alternative value of μ, there is no other unbiased testing procedure that exceeds this one in power – this is uniformly true among all Type I error rates and alternative mean values. By taking the sum of the powers across all tests (i.e., the ETP), I have shown several cases where the estimated ODP does exceed the UMP unbiased test in power, thereby showing that the UMP unbiased test is no longer optimal in the multiple test setting. Furthermore, I have shown that the UMP unbiased test is equivalent to the theoretically optimal procedure only in cases where the ODP thresholding function is symmetric about zero.

To be absolutely rigorous, it should also be shown that the estimated ODP is an unbiased procedure. Otherwise, one can always find a procedure that outperforms the UMP unbiased test for some set of true alternatives even though it may perform badly otherwise. One extension of unbiasedness to multiple significance tests is that for all possible configurations of alternative hypotheses, it holds that $\text{EFP}/m_0 \leq \text{ETP}/(m - m_0)$. It easily follows that the UMP unbiased test satisfies this conditon. The reason for restricting procedures to being unbiased is that one can always define extreme examples that will perform very well in limited situations. For example, we could define a procedure that uses a right-sided significance rule (i.e., a test is significant if $z_i \geq c$). Whenever all true alternative parameters are greater than zero, this procedure will perform as well as the theoretical ODP. However, if enough true alternative parameters are negative, then this procedure will be such that $\text{EFP}/m_0 > \text{ETP}/(m - m_0)$. Thus, when seeking the optimal procedure that can be applied in practice, it makes sense to eliminate such cases.

It has been verified under a number of extreme scenarios that the estimated ODP is unbiased.

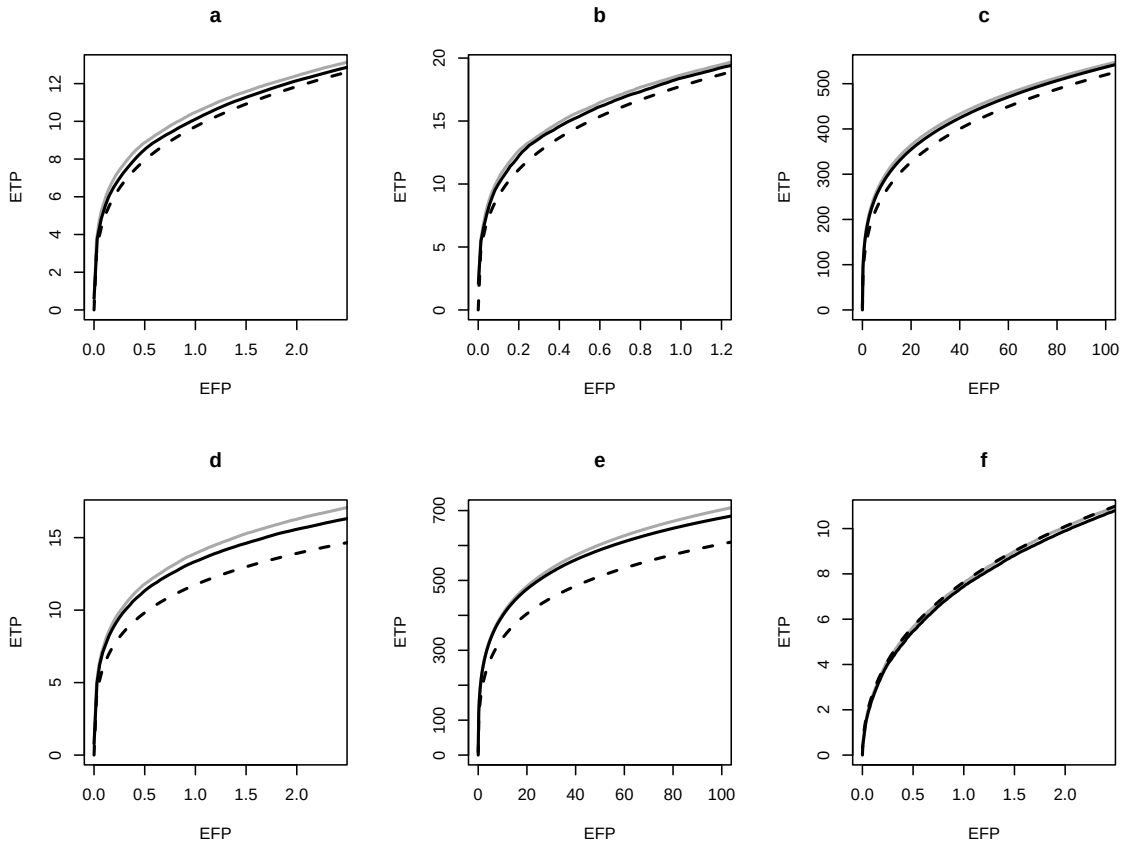

Figure 2: A comparison in terms of ETP versus EFP between the true ODP (grey), estimated ODP (black), and UMP unbiased procedure (dashed). It can be seen that the estimated ODP outperforms the UMP unbiased procedure in all cases except where the signal is perfectly symmetric, in which case the true ODP is equivalent to the UMP unbiased procedure. For each panel, multiple significance tests of $\mu_i = 0$ versus $\mu_i \neq 0$ were performed based on a single observation $z_i \sim N(\mu_i, 1)$ for each test. A number of alternative means were assigned, and there are equal proportions of each one among the false null hypotheses. **(a)** 48 tests, 24 true nulls, alternative means -1, 1, 2, 3. **(b)** 48 tests, 12 true nulls, alternative means -1, 1, 2, 3. **(c)** 2000 tests, 1000 true nulls, alternative means -1, 1, 2, 3. **(d)** 48 tests, 24 true nulls, alternative means 1, 2, 3. **(e)** 2000 tests, 1000 true nulls, alternative means 1, 2, 3. **(f)** 48 tests, 24 true nulls, alternative means -2, -1, 1, 2.

(For example, I made the alternative true for tests 1 and 2, and I set $\mu_1 = 0.01$ and $\mu_2 = -0.01$. The total number of tests was set as $m = 1000$. This is an extremely adverse scenario for the estimated ODP. Even so, it was shown through simulation to be unbiased.) Heuristic arguments also indicate that the estimated ODP is unbiased. This procedure lies somewhere between the extreme of forming a completely symmetric rule, and forming a one-sided adaptive test-specific rule. In the latter case, one would take a right-sided rule if $z_i > 0$, and a left-sided rule if $z_i < 0$. This procedure can easily be shown to be unbiased. Therefore, since the estimated ODP finds a balance between these two extremes, it seems highly unlikely that it would be biased in any circumstance. As stated above, requiring unbiasedness is meant to prevent obviously bad procedures from being considered, so this is likely not an issue for the estimated ODP.

6 Extensions and Connections to Other Concepts

The formulation and optimality of the ODP can be connected to several other well known concepts in statistics. Here I discuss connections to Stein's paradox, shrinkage estimation, FDR, and Bayesian classification.

6.1 Stein's paradox and shrinkage estimation

Stein's paradox (Stein 1956, Stein 1981) had a large impact on notions about high-dimensional point estimation when it was shown that estimating several Normal means can be universally improved by shrinking the usual estimators (the sample means) towards each other – or towards any constant, for that matter. The amount of shrinkage depends on what is in the data as a whole. In other words, the shrunken estimates take into account all of the data at once. This estimator is usually referred to as the James-Stein estimator (James & Stein 1961).

There does not appear to be any previously well established analogue to this paradox in the significance testing setting. However, the formulation of the ODP and the numerical comparisons shown above provide a first step towards this end. There are similarities and differences, however. The results involving the ODP are similar in that it has been shown that borrowing information across tests leads to procedures that are more powerful than optimal single testing procedures. However, we have shown that the estimated ODP does not *always* beat the UMP unbiased procedure (Figure 2f), whereas the James-Stein estimator was shown to always beat the optimal univariate estimator.

Another similarity can be found through the shrinkage estimator interpretation of the James-Stein estimate. As stated above, the James-Stein estimate shrinks each individual sample mean towards a central quantity. Under certain assumptions, the ODP can be written as shrinking the

single test likelihood ratio statistic towards a quantity involving information from the other tests. When the status of each test is made random (that is, each null hypothesis is true with a certain probability), the ODP thresholding rule is defined by

$$\mathcal{S}_{\text{ODP}}(\mathbf{x}) = \frac{g_1(\mathbf{x}) + g_2(\mathbf{x}) + \cdots + g_m(\mathbf{x})}{f_1(\mathbf{x}) + f_2(\mathbf{x}) + \cdots + f_m(\mathbf{x})}.$$

This result is explicitly stated below in Section 6.3, and proved in Section A.2.

The overall ODP statistic can then be written as a weighted average of (i) the individual test's Neyman-Pearson likelihood ratio statistic and (ii) the ODP statistic applied to the remaining tests. For example, hypothesis test 1 has significance thresholding function

$$\mathcal{S}_{\text{ODP}}(\mathbf{x}_1) = \gamma_1 \cdot \frac{g_1(\mathbf{x}_1)}{f_1(\mathbf{x}_1)} + (1 - \gamma_1) \cdot \frac{g_2(\mathbf{x}_1) + \cdots + g_m(\mathbf{x}_1)}{f_2(\mathbf{x}_1) + \cdots + f_m(\mathbf{x}_1)},$$

where the weight γ_1 is specific to test 1. The first term in the weighted sum is the NP statistic applied to test 1, and the second term is the ODP for tests 2 through m applied to test 1. The formula for the weight is $\gamma_1 = f_1(\mathbf{x}_1)/[f_1(\mathbf{x}_1) + f_2(\mathbf{x}_1) + \cdots + f_m(\mathbf{x}_1)]$. The interpretation is that γ_1 quantifies how useful the information from the other tests is: the more similar the null distributions, then the more the ODP uses information from the other tests. This makes sense because it looks at the relative contribution to the EFP from the other tests relative to test 1. Note that when an estimated form of the ODP can be written as above (e.g., the estimate proposed in Storey et al. (2005)), then the interpretation is similar.

6.2 False discovery rate optimality by the ODP

The optimality achieved by the ODP is described in terms of maximizing the ETP for each fixed EFP level. However, it is straightforward to show that this is related to optimality in terms of the FDR, which is currently a popular multiple testing error measure. In what follows I show that minimizing the "missed discovery rate" for each fixed FDR is approximately (and sometimes exactly) equivalent to the optimization that the ODP achieves.

Let FP = false positives, TP = true positives, FN = false negatives, and TN = true negatives. These are the four types of outcomes that occur when applying a significance threshold to multiple significance tests. The FDR is the proportion of false positives among all tests called significant (Soric 1989, Benjamini & Hochberg 1995):

$$\text{FDR} \equiv \text{E}\left[\frac{\text{FP}}{\text{FP} + \text{TP}}\right],$$

where the denominator of the ratio is set to zero when no null hypotheses are rejected. This quantity can be written as a trade-off between the EFP and ETP:

$$\text{FDR} \approx \frac{\text{EFP}}{\text{EFP} + \text{ETP}},$$

where the approximate equality "≈" is sometimes an exactly equality. The approximation applies when testing a large number of hypotheses that are at most weakly dependent so that

$$\lim_{m \to \infty} \left| \text{FDR} - \frac{\text{EFP}}{\text{EFP} + \text{ETP}} \right| = 0.$$

The exact conditions for this convergence have been defined and studied elsewhere (Storey et al. 2004). A variation of the FDR, called positive FDR (pFDR), has been considered that quantifies the proportion of false positives only in cases where at least one test is called significant. Under a Bayesian mixture model assumption (similar to Lemma 3 below), it has been shown that pFDR = EFP/(EFP+ETP), where this is an *exact* equality (Storey 2003). Finally, one can consider another variation on the FDR, which I call the marginal FDR (mFDR), that is simply defined to be the ratio of the EFP to the total number of tests expected to be significant: $\text{mFDR} \equiv \text{EFP}/(\text{EFP} + \text{ETP})$. The important point is that in all of these cases, the FDR can be written and understood in terms of the EFP and the ETP.

It has recently been suggested that FDR optimality should be defined in terms of the proportion of true alternatives among the tests not called significant (Genovese & Wasserman 2002). This quantity has been called the "false non-discovery rate" (Genovese & Wasserman 2002) and the "miss rate" (Taylor et al. 2005); to find a common name, I call it the "missed discovery rate" (MDR). Specifically, a procedure is considered to be optimal if for each fixed FDR level, the MDR is minimized. The above formulations of FDR can easily be extended to the MDR. It can be shown that

$$\text{MDR} \equiv \text{E}\left[\frac{\text{FN}}{\text{FN} + \text{TN}}\right] \approx \frac{\text{EFN}}{\text{EFN} + \text{ETN}},$$

where EFN is expected number of false negatives and ETN is the expected number of true negatives. Again, the approximate equality is sometimes an exact equality. This shows that the MDR can be understood and written in terms of the EFN and ETN. However, the EFN and ETN do not provide any additional information beyond the EFP and ETP:

$$\frac{\text{EFN}}{\text{EFN} + \text{ETN}} = \frac{(m - m_0) - \text{ETP}}{m - \text{ETP} - \text{EFP}}.$$

Therefore, the trade-off between FDR and MDR can be understood in terms of the EFP and the

ETP, which is stated precisely in the following lemma.

Lemma 2. *Suppose that the FDR and MDR are precisely represented in terms of the EFP and ETP as shown above. If the ETP is maximized for each fixed EFP level, then the MDR is minimized for each fixed FDR level. That is, achieving optimality in terms of the EFP and ETP is equivalent to achieving optimality in terms of the FDR and MDR. This implies that the ODP also achieves FDR optimality. When the FDR and MDR may only approximately be written in terms of EFP and ETP, then this equivalence is an approximate one.*

Therefore, the optimality achieved by the ODP is closely related, and sometimes exactly related, to FDR optimality. It can also be argued that EFP and ETP are more fundamental components than the FDR, so perhaps the ODP optimality is more relevant in general.

6.3 ODP under randomized null hypotheses

The ODP was formulated above under the assumption that the status (i.e., truth or falsehood) of each null hypothesis is fixed. Because of this, the ODP is defined in terms of the *true* distribution for each significance test. As it turns out, the status of each test must be modeled as *random* in order for the null and alternative densities of every test to be present in the ODP. The following lemma derives the ODP when the status of each test is randomized.

Lemma 3: Randomized Null ODP. *Suppose that m simple significance tests are performed on observed data sets $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m$, where significance test i has null density f_i and alternative density g_i, for $i = 1, \ldots, m$. Suppose that the each null hypothesis is true with probability π_0. The following significance thresholding function defines the ODP in this case:*

$$\mathcal{S}_{\text{ODP}}(\mathbf{x}) = \frac{g_1(\mathbf{x}) + g_2(\mathbf{x}) + \cdots + g_m(\mathbf{x})}{f_1(\mathbf{x}) + f_2(\mathbf{x}) + \cdots + f_m(\mathbf{x})}.$$

Null hypothesis i is rejected if and only if $\mathcal{S}_{\text{ODP}}(\mathbf{x}_i) \geq \lambda$ for some $0 \leq \lambda < \infty$. For each fixed λ, this procedure yields the maximum ETP among all simultaneous thresholding procedures that have an equal or less EFP, where the expectations are taken over all random quantities.

The proof of this lemma appears in Section A.2; it follows very similarly to that for Lemma 1. Note that no assumptions are made about independence between tests, either among the observed data for each test or the randomized status of the null hypothesis; the lemma holds under arbitrary levels of dependence. This result provides a bridge between the purely frequentist ODP and the Bayesian approach to classification of hypotheses. It also allows us to explore the ODP as a shrunken version of the likelihood ratio test as was presented in the main text.

6.4 Bayesian classification

Besides randomizing the status of the null hypotheses (as was done just above), a prior distribution can be placed on the f_i and the g_i making the null and alternative densities random for each test. Classical Bayesian classification theory begins with these assumptions, which can then be used to find the rule that classifies each test as "true null" or "true alternative" so that the misclassification rate is minimized. As stated earlier, the goal of optimizing ETP for each fixed EFP level is equivalent to minimizing misclassification rate. The optimal Bayesian classifier (called the Bayes rule) has a simple form that can be connected to the NP and ODP approaches. Letting f_B be the expected null density and g_B the expected alternative density, the significance thresholding function can be written as

$$\mathcal{S}(\mathbf{x}) = \frac{g_B(\mathbf{x})}{f_B(\mathbf{x})},$$

where the alternative hypothesis i is classified as true if and only if $\mathcal{S}(\mathbf{x}_i) \geq \lambda$. The choice of λ is driven by the prior distributions and the loss that is placed on each type of error. It should be noted that this is not the usual way for writing the Bayes rule classifier, although it is algebraically equivalent.

From one perspective, this rule is more similar to the Neyman-Pearson approach than to the ODP approach. The reason is that the high-dimensional information across tests is averaged out in forming f_B and g_B, so one is essentially back into the Neyman-Pearson setting. As it turns out, the ODP can be seen as a more specialized version of the Bayes rule, where one instead conditions on the actual realized densities for each test when forming an optimal testing procedure. Since the ODP minimizes the misclassification rate for each realized set of densities, it continues to do so when averaging over all densities. Therefore, the ODP and the Bayes rule both obtain the lowest misclassification rate when averaging over randomized hypotheses and probability density functions.

In the multiple Normal means example, one could put a prior distribution on the μ_i. Suppose that μ_i equals zero (the null case) with probability π_0, and μ_i is drawn from, say, a $N(\theta, 1)$ distribution with probability $1 - \pi_0$. The Bayes rule that minimizes the misclassification rate can be written in terms of the significance thresholding function

$$\mathcal{S}(z) = \frac{\frac{1}{\sqrt{4\pi}} \exp \frac{-(z-\theta)^2}{4}}{\frac{1}{\sqrt{2\pi}} \exp \frac{-z^2}{2}},$$

where one thresholds it as above (Lehmann 1986).

Even though the Bayes rule is more transparently related to the NP, there does exist a direct connection to the ODP. Whereas the above Bayes rule is optimal in terms of misclassification

rate when averaging over all possible outcomes of hypotheses and distributions f_i and g_i, the ODP minimizes misclassification rate for each specific realization of these. Therefore, the ODP also achieves the minimum misclassification rate when averaging over all possible outcomes of hypotheses and distributions, and it can be seen as a more specialized version of the Bayes rule for classification. This is more formally stated in the following lemma.

Lemma 4. *Under the assumptions of Lemma 1, suppose also that the each null hypothesis is true with probability π_0 and that the null and alternative probability density functions, f_i and g_i, for each test are random realizations from their respective prior distributions. The ODP of Lemma 1 achieves the minimum misclassification rate when applied to any specific realization of null and alternative hypotheses and densities. This is true both conditional on the specific realizations and when averaging over the entire population of configurations of the hypotheses, null densities and alternative densities. The latter fact implies that it also achieves the minimum misclassification rate in the Bayesian setting.*

The proof of this lemma is straightforward and is outlined as follows. Suppose that Type I errors are given relative weight w and Type II errors weight $1-w$. The misclassification rate for a fixed set of true null hypotheses and densities is then $w\text{EFP} + (1-w)[(m-m_0) - \text{ETP}]$. It is straightforward to show that because the ODP maximizes ETP for each fixed EFP, the misclassification rate is also minimized for some EFP level. That is, for each fixed w and m_0 the minimum misclassification rate requires a specific EFP level (that can be calculated directly), which in turn specifies the appropriate cut-off to apply to the ODP thresholding function. (Keep in mind that EFP and ETP are expectations conditional on specific realizations of null and alternative hypotheses and densities.) If the ODP is optimal for each realization, then it will remain so when averaging over all possible realizations of true null hypotheses and probability densities.

A question remains as to whether it is better to apply the ODP or the Bayes rule when doing a large number of tests. Although quite difficult to answer in general, we have shown that an estimated ODP substantially outperforms several empirical Bayesian methods when attempting to identify differentially expressed genes in DNA microarray experiments (Storey et al. 2005). One explanation is that the Bayesian approach derives its optimal rule by averaging over a set of outcomes much larger than the one at hand, whereas the ODP is directed at the set of null and alternative distributions present in that one experiment. Another explanation is that an estimated ODP makes less assumptions, thereby outperforming the empirical Bayesian methods. Indeed, the estimated ODP proposed in Storey et al. (2005) is shown to have an empirical Bayesian interpretation that indicates very few assumptions are being made about the prior distributions.

7 Discussion

A new statistical theory has been developed here that shows how to optimally perform multiple significance tests based on a simultaneous thresholding rule. The ODP allows one to simultaneously test multiple hypotheses in such a way that the total number of expected true positives is maximized for each fixed number of expected false positives. This procedure can be viewed as a multiple test extension of the Neyman-Pearson procedure for testing a single hypothesis. The ODP has connections to several different areas of statistics, including FDRs, Bayesian classification, shrinkage estimation, and Stein's paradox.

The recent explosion in research on multiple testing has consistently started with the assumption that p-values are obtained for each test individually. In contrast, the ODP is a truly high-dimensional approach that uses all of the relevant information across tests when assessing the significance of each one. The ODP method does not merely modify existing single test procedures, thus apparently representing a significant departure from the current point of view.

I have briefly discussed a strategy for implementing the ODP in practice. However, this is not an easy task, and it will take some substantial developments to arrive at a generally applicable set of methods. We have applied and extended the ideas here to applications in genomics to produce an estimated version of the ODP for identifying genes that are differentially expressed in comparative microarray experiments (Storey et al. 2005). This method shows surprisingly strong gains in power relative to a number of leading methods currently available. It is also pointed out there that the ODP strategy may be useful in a variety of high-dimensional biological studies due to the fact that there is often a strong and largely unknown structure among the significance tests, where the goal is to extract as much biologically meaningful signal as possible.

A Appendix

A.1 Ubiquity of simultaneous thresholding rules

The ODP is optimal among all procedures that apply a *single* simultaneous thresholding rule to each test. Even though it is possible to consider many different arrangements of thresholding functions that could be used to define optimality, the simultaneous thresholding function is typically the only one available in practice. This is the main motivation for the formulation of the ODP in terms of a simultaneous thresholding rule. This is further motivated by the fact that an *estimate* of ODP may substantially outperform an *estimate* of a test-specific optimal thresholding rule (see the comparison of the estimated ODP to the UMP unbiased test below, as well as our comparison of the estimated ODP to a generalized likelihood ratio test in Storey et al. (2005)).

In the multiple Normal means example above, the UMP unbiased procedure invokes a simultaneous thresholding rule when applied to several tests. This is easily shown to be true in other standard situations. For example, suppose that a standard two-sided t-test is applied to each $\mathbf{x}_i$. The statistic is

$$\mathcal{S}(\mathbf{x}_i) = \left| \frac{\overline{x}_i}{s_i/\sqrt{n}} \right|,$$

where $\overline{x}_i$ is the sample mean, and s_i is the sample standard deviation of $\mathbf{x}_i$. Since the sample mean $\overline{x}_i$ and standard error $s_i/\sqrt{n}$ are functions of $x_{i1}, \ldots, x_{in}$, it follows that $\mathcal{S}(\mathbf{x}_i)$ is also a function of these data. Also, each test has the same null distribution. Therefore, test i is called significant if and only if $\mathcal{S}(\mathbf{x}_i) \geq \lambda$, making the standard two-sided t-test a simultaneous thresholding rule when applied to multiple tests. The following lemma states in general terms when multiple significance testing procedures are equivalent to simultaneous thresholding rules.

Lemma 5. *Suppose multiple significance tests are performed where (i) the data for each test follow a common family of distributions, where all possible differences in parameters are unknown, (ii) the null and alternative hypotheses are identically defined for each test, (iii) the same procedure is applied to each test in estimating unknown parameter values. Any differences between test statistics are then due to differences in their data, not prior knowledge about the significance tests. In this case, the relative significance among the tests is based only on the data from each test, implying that for any thresholding rule there is an equivalent simultaneous thresholding rule.*

It is easily verified that Lemma 5 describes the typical multiple testing scenarios encountered in practice. Exceptions occur when prior information about the tests is known that distinguishes them. In such a case, one can nevertheless arrange the tests into the largest groups possible that meet the above criteria and apply the ODP to each set individually. The most extreme version of this is when the exact distributions are *known* for each test, and these distributions differ from test to test. In this case, each test is its own largest group, and the NP procedure is applied to each one individually. However, this is quite different than if one is merely *aware* that the distributions differ from test to test; in this case, Lemma 5 is likely to apply because the distributions would have to be estimated by some common data based procedure.

A.2 Proofs of Lemmas

The proofs of Lemmas 1, 3 and 5 follow. The proofs of Lemmas 2 and 4 are clear from the text.

Proof of Lemma 1. Let $\Gamma_\lambda = \{\mathbf{x} : \mathcal{S}_{\text{ODP}}(\mathbf{x}) \geq \lambda\}$ be the significance region for the ODP applied

at cut-off λ. The EFP and ETP of any general significance region Γ are:

$$\begin{aligned}\text{EFP}(\Gamma) &= \int_\Gamma f_1(\mathbf{x})d\mathbf{x} + \cdots + \int_\Gamma f_{m_0}(\mathbf{x})d\mathbf{x},\\ \text{ETP}(\Gamma) &= \int_\Gamma g_{m_0+1}(\mathbf{x})d\mathbf{x} + \cdots + \int_\Gamma g_m(\mathbf{x})d\mathbf{x}.\end{aligned}$$

The goal is then to show that for any Γ' such that $\text{EFP}(\Gamma') \leq \text{EFP}(\Gamma_\lambda)$, it is the case that $\text{ETP}(\Gamma') \leq \text{ETP}(\Gamma_\lambda)$. Proving this optimality at first seems difficult because it has to be shown over m different random variables. However, a calculation trick greatly simplifies achieving this goal.

Note that it is equivalent to show that $\text{EFP}(\Gamma')/m_0 \leq \text{EFP}(\Gamma_\lambda)/m_0$ implies $\text{ETP}(\Gamma')/(m-m_0) \leq \text{ETP}(\Gamma_\lambda)/(m-m_0)$. To this end, define $\overline{f} = \left[\sum_{i=1}^{m_0} f_i\right]/m_0$ and $\overline{g} = \left[\sum_{i=m_0+1}^{m} g_i\right]/(m-m_0)$; it is easily verified that these function each integrate to one. It then follows that

$$\begin{aligned}\frac{\text{EFP}(\Gamma)}{m_0} &= \frac{\int_\Gamma f_1(\mathbf{x})d\mathbf{x} + \cdots + \int_\Gamma f_{m_0}(\mathbf{x})d\mathbf{x}}{m_0}\\ &= \int_\Gamma \frac{f_1(\mathbf{x}) + \cdots f_{m_0}(\mathbf{x})}{m_0}d\mathbf{x} = \int_\Gamma \overline{f}(\mathbf{x})d\mathbf{x},\\ \frac{\text{ETP}(\Gamma)}{m-m_0} &= \int_\Gamma \overline{g}(\mathbf{x})d\mathbf{x}.\end{aligned}$$

It should be pointed out that there is no frequentist probabilistic intepretation of $\overline{f}$ and $\overline{g}$ used in this proof. Since these functions each integrate to one, the Neyman-Pearson lemma can be invoked as a mathematical optimization result among this class of functions. According to the Neyman-Pearson lemma, the significance regions based on $\overline{g}(\mathbf{x})/\overline{f}(\mathbf{x})$ are optimal for maximizing $\int_\Gamma \overline{g}(\mathbf{x})d\mathbf{x}$ for each fixed level of $\int_\Gamma \overline{f}(\mathbf{x})d\mathbf{x}$. However,

$$\mathcal{S}_{\text{ODP}}(\mathbf{x}) = \frac{(m-m_0)\overline{g}(\mathbf{x})}{m_0\overline{f}(\mathbf{x})},$$

making the two thresholding functions equivalent. Therefore, the ODP maximizes ETP for each fixed EFP.

Proof of Lemma 3. The proof is similar to that for Lemma 1. Let $\pi_1 = 1 - \pi_0$, where π_0 is the prior probability that any null hypothesis is true. The EFP and ETP for any general significance region Γ are:

$$\begin{aligned}\text{EFP}(\Gamma) &= \int_\Gamma \pi_0 f_1(\mathbf{x})d\mathbf{x} + \cdots + \int_\Gamma \pi_0 f_m(\mathbf{x})d\mathbf{x},\\ \text{ETP}(\Gamma) &= \int_\Gamma \pi_1 g_1(\mathbf{x})d\mathbf{x} + \cdots + \int_\Gamma \pi_1 g_m(\mathbf{x})d\mathbf{x}.\end{aligned}$$

The proof follows exactly as above, except that $\overline{f} = \left[\sum_{i=1}^{m} f_i\right]/m$ and $\overline{g} = \left[\sum_{i=1}^{m} g_i\right]/m$. Note that this proof is easily extended to accommodate both (i) priors that differ from test to test and (ii) versions of the EFP and ETP where each test may be weighted differently.

Proof of Lemma 5. Suppose that under the assumptions of Lemma 5, it is possible to employ two thresholding rules. Therefore, there exists significance tests i and j, such that $\mathcal{S}_1$ is applied to i and $\mathcal{S}_2$ is applied to j. According to the assumptions of the lemma, it is not possible to distinguish these tests before observing any data. Therefore the assignments of tests i and j to their respective thresholding functions is based on observed data. Because of this, one may define an indicator function Δ such that if $\Delta(\mathbf{x}) = 1$, then thresholding function $\mathcal{S}_1$ is used, and if $\Delta(\mathbf{x}) = 0$, then thresholding function $\mathcal{S}_2$ is used. (Note that Δ, $\mathcal{S}_1$ and $\mathcal{S}_2$ could be data dependent functions, but these must be formed from looking at all of the data, making them eventually definable as functions of each test's individual data set.) In this case the simultaneous thresholding function $\mathcal{S}(\mathbf{x}) = \Delta(\mathbf{x})\mathcal{S}_1(\mathbf{x}) + [1 - \Delta(\mathbf{x})]\mathcal{S}_2(\mathbf{x})$ is equivalent, leading to a contradiction. A similar argument follows if one begins with the assumption that it is possible to apply K different thresholding functions.

Acknowledgments

This research was supported in part by NIH grant R01 HG002913-01. Thanks to Alan Dabney for a number of useful comments on the manuscript.

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.

Genovese, C. & Wasserman, L. (2002). Operating characteristics and extensions of the FDR procedure, *Journal of the Royal Statistical Society, Series B* **64**: 499–517.

James, W. & Stein, C. (1961). Estimation with quadratic loss, *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability* **1**: 361–379.

Lehmann, E. L. (1986). *Testing Statistical Hypotheses*, second edn, Springer-Verlag.

Neyman, J. & Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses, *Philosophical Transactions of the Royal Society* **231**: 289–337.

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

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

Stein, C. (1956). Inadmissability of the usual estimator for the mean of a multivariate distribution, *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability* **1**: 197–206.

Stein, C. (1981). Estimation of the mean of a multivariate Normal distribution, *Annals of Statistics* **9**: 1135–1151.

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., Dai, J. Y. & Leek, J. T. (2005). Optimal separation of signal from noise in high-dimensional biological studies. *UW Biostatistics Working Paper Series*, Working Paper 260. `http://www.bepress.com/uwbiostat/paper260/`.

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.

Taylor, J., Tibshirani, R. & Efron, B. (2005). The miss rate for the analysis of gene expression data, *Biostatistics* **6**: 111–117.