Multiple hypothesis testing using the excess discovery count and alpha-investing rules

Dean P. Foster and Robert A. Stine*
Department of Statistics
The Wharton School of the University of Pennsylvania
Philadelphia, PA 19104-6340

April 7, 2005

Abstract

We propose an adaptive, sequential methodology for testing multiple hypotheses. Our methodology consists of a new criterion, the excess discovery count (EDC), and a new class of testing procedures that we call alpha-investing rules. The excess discovery count is the difference between the number of correctly rejected null hypotheses and a fraction of the total number of rejected hypotheses. EDC shares many properties with the false discovery rate (FDR), but is adapted to testing a sequence of hypotheses rather than a fixed set. Because EDC controls the count of incorrectly rejected hypotheses rather than a ratio, we are able to prove that a wide class of testing procedures that we call alpha-investing rules control EDC. Alpha-investing rules mimic alpha-spending rules used in sequential trials, but possess a key difference. When a test rejects a null hypothesis, alpha-investing rules earn additional probability toward testing subsequent hypotheses. Alpha-investing rules allow one to incorporate domain knowledge into the testing procedure and improve the power of the tests.

Key words and phrases: Bonferroni method, false discovery rate (FDR), family wide error rate (FWER), multiple comparison procedure.

*All correspondence regarding this manuscript should be directed to Prof. Stine at the address shown with the title. He can be reached via e-mail at stine@wharton.upenn.edu.

1 Introduction

We propose an adaptive, sequential methodology for testing multiple hypotheses. Our approach works in the usual setting in which one has a batch of several hypotheses as well as cases in which the hypotheses arrive sequentially in a stream. Streams of hypothesis tests arise naturally in variety of contemporary modeling applications, such as genomics and variable selection for large models. In contrast to the comparatively well-defined problems that spawned multiple comparison procedures such as Tukey's studentized range, these applications can involve thousands of tests. For example, microarrays lead one to compare a control group to a treatment group using measured differences on over 6,000 genes (Dudoit, Shaffer and Boldrick, 2003). In contrast, the example used by Tukey to motivate the problems of multiple comparisons compares the means of only 6 groups (Tukey, 1953, available in Braun (1994)). If one considers the possibility for interactions, then the number of tests is virtually infinite. Because our approach allows the testing to proceed sequentially, the choice of future hypotheses can depend upon the results of previous tests. Thus, having discovered differences in certain genes, an investigator could, for example, direct attention toward related genes identified by common transcription factor binding sites (Gupta and Ibrahim, 2005).

Our methodology has two key components, a criterion and a procedure. For multiple testing, we distinguish criteria that control the number of Type I errors from testing procedures. We call our new criterion the *excess discovery count* (EDC). EDC tracks the expected number of true rejections among the rejected hypotheses. To control EDC, a test procedure must guarantee that the expected count of true rejections exceeds a chosen fraction of the number of rejected hypotheses. For example, one might want to guarantee that at least 95% of the rejected hypotheses were rejected correctly. Although one can use EDC to control traditional tests, the advantage of this criterion is that it permits one to control adaptive testing procedures in which the choice of the next hypothesis to test depends on previous results.

The second component of our methodology is a class of adaptive testing procedures that we call *alpha-investing rules.* We show that testing procedures in this class control EDC. Alpha-investing rules allow one to test a possibly infinite stream of hypotheses, accommodate dependent tests, and incorporate domain knowledge. Alpha-investing rules mimic alpha-spending rules that are commonly used in clinical trials. Unlike alpha-spending rules, however, alpha-investing rules treat each test as an "investment."

Each test has a cost, but can generate a profit in the form of the an increase in the amount of Type I error available for subsequent tests.

The rest of this paper develops as follows. We first review several ideas from the literature on multiple comparisons, particularly those related to the family wide error rate and the false discovery rate. With these ideas in place, we define EDC in Section 3 and alpha-investing rules in Section 4. In Section 5, we show that alpha-investing rules control a generalized version of EDC. We give several examples of testing a sequence of hypotheses using alpha-investing rules in Section 6. We close in Section 7 with a brief summary discussion, and defer the single proof to the appendix.

2 Criteria and Procedures

We begin with a brief review of criteria and procedures used to test a collection of hypotheses. To set the stage for describing EDC, we review the two most important criteria commonly applied in testing multiple hypotheses: the family wide error rate and the false discovery rate. These criteria generalize the notion of the Type I error rate (α-level) to tests of several hypotheses and are often confused with testing procedures. The false discovery rate is a criterion that one might design a testing procedure to satisfy, but is not itself a testing procedure. Just as there are many α-level tests of a simple hypothesis, so too are there various multiple testing procedures. We confine our attention to two, the Bonferroni procedure and step-up/step-down tests. These procedures are most closely related to and suggestive of the alpha-investing rules developed in Section 4.

Suppose that we have a set of m null hypotheses $\mathcal{H}(m) = \{H_1, H_2, \ldots, H_m\}$ that specify values for parameters $\theta = \{\theta_1, \theta_2, \ldots, \theta_m\}$. Each parameter θ_j can be scalar or vector-valued, and Θ denotes the space of parameter values. In the most familiar case, each null hypothesis specifies that a parameter is zero, $H_j : \theta_j = 0$. We describe the situation in which every hypothesis has this form and is true as the "null model."

We follow the standard notation for labeling the true and false rejections as shown in Table 1, which is taken from Benjamini and Hochberg (1995). Assume that m_0 of the null hypotheses in $\mathcal{H}(m)$ are true. The *observable* statistic $R(m)$ counts how many of these m hypotheses are rejected. The *unobservable* random variable $V^\theta(m)$ denotes the number of false positives among the m tests, those cases in which the testing procedure incorrectly rejects a true null hypothesis. Similarly, $S^\theta(m) = R(m) - V^\theta(m)$ counts the

Table 1: *Counts of the number of null hypotheses that are true and false, displayed as sums of unobserved random variables. The marginal random variable $R(m)$ that counts the total number rejected is observable, but internal counts such as $V^\theta(m)$ depend upon θ.*

		Claim		
		Accept H_0	Reject H_0	
True	H_0	$U^\theta(m)$	$V^\theta(m)$	m_0
State	H_0^c	$T^\theta(m)$	$S^\theta(m)$	$m - m_0$
		$m - R(m)$	$R(m)$	m

number of correctly rejected null hypotheses. We index these random variables with a superscript θ to distinguish them from a statistic such as $R(m)$; $V^\theta(m)$ and $S^\theta(m)$ are not observable without θ. For a null model, $m_0 = m$, $V^\theta(m) = R(m)$ and $S^\theta(m) = 0$.

A basic premise of multiple testing is to control the chance for *any* false rejection. The *family wide error rate* (FWER) is the probability of falsely rejecting *any* null hypothesis from $\mathcal{H}(m)$, regardless of the values of the underlying parameters,

$$\text{FWER}(m) \equiv \sup_{\theta \in \Theta} \mathrm{P}_\theta(V^\theta(m) \geq 1) \ . \tag{1}$$

An important special case is control of FWER under the null model. We refer to this criterion as the *size* of a procedure,

$$\text{Size}(m) = \mathrm{P}_0(V^\theta(m) \geq 1) \ , \tag{2}$$

where P_0 denotes the probability measure under the null model. All of the procedures that we describe control Size(m), but not all control the more general FWER.

The Bonferroni procedure is familiar and represents an important benchmark for comparison. Let $p_1, \ldots, p_m$ denote the p-values of tests of $H_1, \ldots, H_m$. Given a chosen level $0 < \alpha < 1$, the usual Bonferroni procedure rejects those H_j for which $p_j \leq \alpha/m$. Let the indicators $V_j^\theta \in \{0, 1\}$ track incorrect rejections; $V_j^\theta = 1$ if H_j is incorrectly rejected and is zero otherwise. Then $V^\theta(m) = \sum V_j^\theta$ and the inequality

$$\mathrm{P}_\theta(V^\theta(m) \geq 1) \leq \sum_{j=1}^{m} \mathrm{P}_\theta(V_j^\theta = 1) \leq \alpha \tag{3}$$

shows that this procedure controls FWER$(m) \leq \alpha$. More generally, one need not distribute α equally over $\mathcal{H}(m)$; the procedure only requires that the sum of the α-levels is not more than α. For example, alpha-spending rules allocate α over a collection of hypotheses with a larger share given to hypotheses of greater interest. Although it controls FWER, the Bonferroni procedure is often criticized for having little power compared to other methods. Clearly, its power decreases as m increases because the threshold α/m for detecting a significant effect decreases.

To obtain more power when some null hypotheses are false but still control FWER, Holm (1979) introduced the following so-called step-down testing procedure. Order the collection of m hypotheses so that the p-values of the associated test statistics are sorted from smallest to largest (putting the most significant first),

$$p_{(1)} \le p_{(2)} \le \cdots \le p_{(m)} .$$

The test of $H_{(1)}$ has p-value $p_{(1)}$, the test of $H_{(2)}$ has p-value $p_{(2)}$ and so forth. Holm's procedure rejects those hypotheses $H_{(j)}$ for which $p_{(j)}$ is less than an *increasing* sequence of thresholds. The procedure first compares the smallest p-value to the Bonferroni threshold. If $p_{(1)} > \alpha/m$, the procedure stops and does not reject any hypothesis. Consequently, $\text{Size}(m) \le \alpha$. If $p_{(1)} \le \alpha/m$, the procedure rejects $H_{(1)}$ and moves on to test $H_{(2)}$. Rather than compare $p_{(2)}$ to α/m, however, Holm's procedure compares $p_{(2)}$ to a larger threshold, $\alpha/(m-1)$. In general, if we define $j_d = \min\{j : p_{(j)} > \alpha/(m-j+1)\}$, then Holm's step-down procedure rejects $H_{(1)}, \ldots, H_{(j_d-1)}$. Because of the nesting, this testing procedure is closed in the sense of Marcus, Peritz and Gabriel (1976) and hence controls $\text{FWER}(m) \le \alpha$. Obviously, when compared to using the Bonferroni threshold for each p-value, Holm's method has larger power. The improvement is small, however, when m is large because α/m is so close to $\alpha/(m-j)$ when testing the smallest p-values.

The *false discovery rate* (FDR) criterion controls the size of a testing procedure but introduces a different type of control if the null model is rejected. Benjamini and Hochberg (1995) define FDR as the expected proportion of false positives among rejected hypotheses,

$$\text{FDR}(m) = E_\theta \left(\frac{V^\theta(m)}{R(m)} \mid R(m) > 0 \right) \text{P}(R(m) > 0) . \qquad (4)$$

For the null model, $R(m) = V^\theta(m)$ and $\text{FDR}(m) = \text{FWER}(m)$. Thus, test procedures that control $\text{FDR}(m) \le \alpha$ have $\text{Size}(m) \le \alpha$. Under the alternative, $\text{FDR}(m)$ decreases as the number of false null hypotheses $m - m_0$ increases (Dudoit et al., 2003). As a result, $\text{FDR}(m)$ becomes more easy to control in the presence of non-zero effects, allowing more powerful procedures. Variations on FDR include pFDR (which drops the term $\mathrm{P}(R > 0)$ Storey, 2002, 2003) and the local false discovery rate $\mathrm{fdr}(z)$ (which estimates the false discovery rate as a function of the size of the test statistic Efron, 2005a,b). Closer to our work, Meinshausen and Rice (2004) and Meinshausen and Buehlmann (2004) consider estimates of m_0, the total number of false hull hypotheses in $\mathcal{H}(m)$.

Benjamini and Hochberg (1995) show that the following so-called step-up testing procedure controls FDR. First, assume that the p-values are independent and define $j_u^* = \max\{j : p_{(j)} \le j\,\alpha/m\}$. Using the inequality of Simes (1986), they show that the testing procedure that rejects $H_{(1)}, \ldots, H_{(j^*)}$ controls $\mathrm{FDR}(m) \le \alpha$. This testing procedure thus controls $\mathrm{Size}(m) \le \alpha$, but does not control FWER for all θ. A similar step-down procedure that rejects $H_{(1)}, \ldots, H_{(j_d^*-1)}$ for $j_d^* = \min\{j : p_{(j)} > \alpha/(m-j+1)\}$ also has $\mathrm{FDR}(m) \le \alpha$. Although this step-down procedure has less power than its step-up cousin (because $j_d^* - 1 \le j_u^*$), it has more power than Holm's procedure. Holm's step-down procedure sets thresholds for the p-values to $\frac{\alpha}{m}, \frac{\alpha}{m-1}, \frac{\alpha}{m-2}, \ldots$ whereas a Simes-based step-down procedure uses the larger thresholds $\frac{\alpha}{m}, \frac{2\alpha}{m}, \frac{3\alpha}{m}, \ldots$. A cost of this greater power is a restriction to independent tests that Holm's procedure does not require. Subsequent papers (such as Benjamini and Yekutieli, 2001; Sarkar, 1998; Troendle, 1996) consider situations in which this type of step-up/step-down testing controls FDR under dependence, but the results obtain only for certain types of dependence.

3 The Excess Discovery Count (EDC)

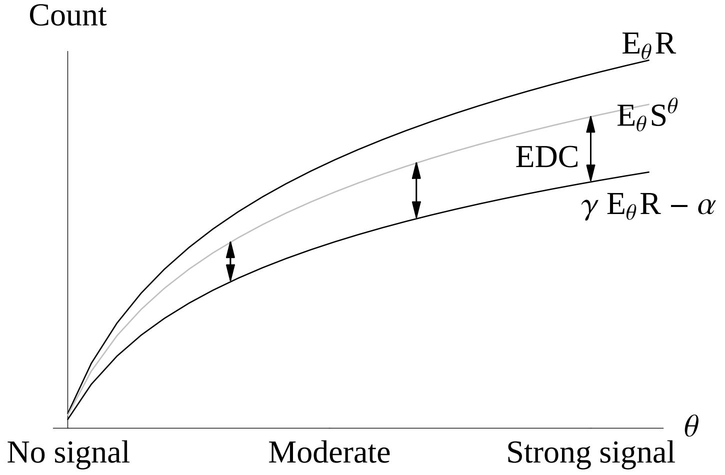
The excess discovery count (EDC) is a new criterion for controlling a multiple testing procedure. Its form resembles that of FDR, and it too controls an unobservable random variable. EDC operates in the domain of counts, however, rather than ratios of counts, and EDC emphasizes the number of correct rejections $S^\theta(m)$ rather than the number of incorrect rejections $V^\theta(m)$. EDC is the expected difference between the number of correctly rejected null hypotheses $S^\theta(m)$ and a fraction $0 \le \gamma \le 1$ of the number of rejected hypotheses $R(m)$ (see Figure 1). For a procedure that tests $\mathcal{H}(m)$, we have

Definition 1. The excess discovery count criterion for testing a set of m hypotheses is

$$\mathrm{EDC}_{\alpha,\gamma}(m) = E_\theta[S^\theta(m) - \gamma\, R(m)] + \alpha\ , \quad 0 < \alpha, \gamma < 1\ . \tag{5}$$

Typical values for the two tuning parameters α and γ are 0.05 and 0.95, respectively.

Figure 1: *EDC controls the gap between the number of true rejections S^θ and a fraction of the number of rejected null hypotheses. A strong signal implies most of the null hypotheses in $\mathcal{H}$ are false.*



FDR(m) controls the expected *proportion* of false positives $V^\theta(m)/R(m)$ given that $R(m) > 0$. EDC$_{\alpha,\gamma}(m)$ instead controls the expected *difference* in the counts $S^\theta(m) - \gamma\ R(m)$. Being a ratio, $0 \leq$ FDR$(m) \leq 1$ and hence resembles a conditional probability. In contrast EDC$_{\alpha,\gamma}(m)$ need not be positive, let alone lie between 0 and 1.

We are most interested in procedures such as that suggested by Figure 1 for which EDC is positive. In this figure, the x-axis indicates the amount of signal in the sense of the proportion of null hypotheses in $\mathcal{H}$ that are false. “Strong signal” implies that many of the m hypothesis are false, whereas “no signal” implies the null model. We will say that a multiple testing procedure “controls EDC” if EDC$_{\alpha,\gamma}(m) \geq 0$. Control of EDC amounts to showing that the expected count of true rejections is at least $\gamma E_\theta R(m) - \alpha$. Under the null model, $S^\theta(m) = 0$ so that

$$\mathrm{EDC}_{\alpha,\gamma}(m) = \alpha - \gamma\, E_\theta R(m) \leq \alpha - \gamma\, \mathrm{Size}(m)\ .$$

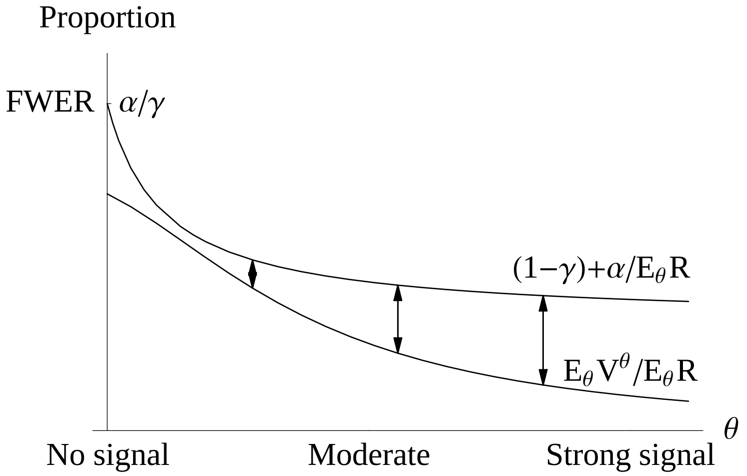
Thus, a procedure that controls EDC$_{\alpha,\gamma}(m) \geq 0$ also controls Size$(m) \leq \alpha/\gamma$. One can also use EDC to control FWER. If $\gamma = 1$, control of EDC implies control of FWER because

$$\mathrm{EDC}_{\alpha,1}(m) \geq 0 \quad \Rightarrow \quad \mathrm{P}_\theta(V^\theta(m) \geq 1) \leq E_\theta V^\theta(m) \leq \alpha\ .$$

This property suggests that one can think of α as controlling the FWER when $\gamma \approx 1$.

The second tuning parameter γ more closely resembles FDR in the sense of controlling the procedure once it rejects the null model. Assuming that $E_\theta\, R(m) > 0$, control

Figure 2: *When viewed as controlling the proportion of false positives among rejected null hypotheses, EDC controls the gap between the ratio of expectations EV^θ/ER and a decreasing function of the number of rejected null hypotheses. A strong signal in the heuristic sense here implies most of the null hypotheses in $\mathcal{H}$ are false.*



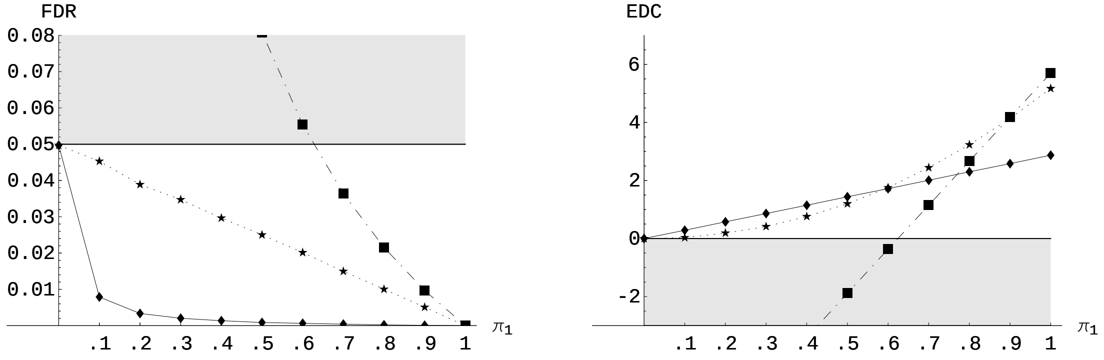
of $\text{EDC}_{\alpha,\gamma}(m)$ implies

$$E_\theta[S^\theta(m) - \gamma\ R(m)] + \alpha \geq 0 \qquad \Rightarrow \qquad \frac{E_\theta\ V^\theta(m)}{E_\theta\ R(m)} \leq (1-\gamma) + \frac{\alpha}{E_\theta R(m)}$$

When many hypotheses in $\mathcal{H}(m)$ are false and $R(m)$ is large, most of the control on the procedure comes from γ. Figure 2 shows EDC from this "FDR point of view" that emphasizes the ratio $E_\theta V^\theta(m)/E_\theta R(m)$ rather than counts. The FDR criterion on this scale is a horizontal line anchored at FWER that controls $E_\theta\left(V^\theta(m)/R(m)\right)$ rather than the ratio of expectations. A criterion that controls the ratio of expectations (rather than the expectation of the ratio) has been discussed in Benjamini and Hochberg (1995).

To supplement these sketches, we ran a small simulation. Figure 3 shows simulated values of FDR and EDC for testing a collection of $m = 200$ hypotheses using three procedures: a naive, fixed-level test that rejects H_j if $p_j \leq \alpha = 0.05$, the step-down Simes procedure, and the standard Bonferroni procedure. The tested hypotheses H_j : $\mu_j = 0$ specify the means of 200 normal populations. We set the values of the μ_j by sampling a spike-and-slab mixture. The mixture puts $100(1-\pi_1)\%$ of its probability in a spike at zero; $\pi_1 = 0$ identifies the null model. The slab of this mixture – the

Figure 3: *FDR (left) and EDC (right, with $\alpha = 0.05$ and $\gamma = 0.95$) control the size of test procedures ($\pi_1 = 0$) and the number rejected as the level of signal π_1 grows. The lines show FDR and EDC for the Bonferroni procedure (—), Simes-based step-down testing ($\cdots$), and a naive procedure that rejects each hypothesis at level $\alpha = 0.05$ ($- \cdot -$).*



signal – is a normal distribution, so that

$$\mu_j \sim \begin{cases} 0 & w.p. \quad 1-\pi_1 \\ N(0,\sigma^2) & w.p. \quad \pi_1 \end{cases} . \tag{6}$$

We set the variance of the signal component of the mixture to $\sigma^2 = 2\log m$ so that the standard deviation of the non-zero μ_j matches the bound commonly used in hard thresholding. The test statistics are independent, normally distributed random variables $Z_j \overset{\text{iid}}{\sim} N(\mu_j, 1)$ for which the two-sided p-values are $p_j = 2(1 - \Phi(|Z_j|))$. Given these p-values, we computed FDR and $\text{EDC}_{0.05,0.95}$ in a simulation with 10,000 trials. In the simulation, we varied the amount of signal varying π_1 from 0 (the null model) to 1.

Qualitatively, FDR and EDC perform similarly. The shaded regions in Figure 3 indicate lack of control of the indicated criterion. Bonferroni and step-down testing control FWER(200) ≤ 0.05 and $\text{EDC}_{\alpha,\gamma}(200) \geq 0$. Simulated values of these criteria remain outside of the shaded regions for all values of π_1. On the other hand, the naive procedure that tests all 200 hypotheses at level 0.05 produces results that fall into the shaded region for many values of π_1. Both FDR and EDC show this procedure as switching from liberal (shaded region) to conservative at about the same level of signal, namely $0.6 < \pi_1 < 0.7$. Notice that FDR emphasizes, relatively speaking, differences among the procedures when the amount of signal is small; as π_1 nears 1, FDR falls to zero for all 3 procedures. Dudoit et al. (2003) discuss this aspect of FDR further. EDC preserves a more uniform scale for various amounts of signal. We note also that the Bonferroni procedure produces linear trends in EDC. The slope of the line seen in the right panel of Figure 3 depends upon the choice of γ in $\mathrm{EDC}_{\alpha,\gamma}$. Conservative methods force $V^\theta(m)$ to be small regardless of the presence of signal so that $E_\theta\left(S^\theta(m) - \gamma\, R(m)\right) + \alpha \approx c(1-\gamma)\,\pi_1$.

4 Alpha-Investing Rules

Alpha-investing rules provide a framework for devising multiple testing procedures that control EDC in a dynamic setting that allows streams of hypotheses. Alpha-investing rules resemble alpha-spending rules such as those often used in sequential clinical trials. In a sequential trial, investigators routinely monitor the accumulating results for safety and efficacy. This monitoring leads to a sequence of tests of one (or several) null hypotheses as the data accumulate. Alpha-spending (or error-spending) rules control the level of such tests. Given an overall Type I error rate for the trial, such as $\alpha = 0.05$, alpha-spending rules allocate, or spend, α over a sequence of tests. As Tukey (1991) writes, “Once we have spent this error rate, it is gone.” When repeatedly testing one null hypothesis H_0 in a clinical trial, spending rules guarantee that $P(\text{reject } H_0) \leq \alpha$ when H_0 is true.

While similar in that they allocate Type I error over multiple tests, alpha-investing rules differ from alpha-spending rules in the following way. An alpha-investing rule earns additional probability toward subsequent Type I errors with each rejected hypothesis. Rather than treating each test as an expense that consumes its Type I error rate, an alpha-investing rule treats tests as investments, motivating our choice of name. In keeping with this analogy, we call the Type I error rate available to the rule its alpha-wealth. As with an alpha-spending rule, an alpha-investing rule can never spend more than its current alpha-wealth. Unlike an alpha-spending rule, however, an alpha-investing rule earns an increment in its alpha-wealth each time that it rejects a null hypothesis. For alpha-investing, Tukey’s remark becomes “If we invest the error rate wisely, we’ll earn more for further tests.” A procedure that invests its alpha-wealth in testing hypotheses that are rejected accumulates additional wealth toward subsequent tests. The more hypotheses that are rejected, the more alpha-wealth it earns. If the test of H_j is not significant, however, the rule loses the α-level invested in this test and its alpha-wealth decreases. The more wealth a rule invests in testing hypotheses that are not rejected, the less alpha-wealth remains for subsequent tests.

More specifically, an alpha-investing rule is a function $\mathcal{I}$ that determines the α-level for testing the next hypothesis in a sequence of tests. We assume an exogenous system external to the investing rule determines the next hypothesis to test. (Though not part of the investing rule itself, this exogenous system can use the sequence of rejections R_j to determine the next hypothesis to test.) An alpha-investing rule has two parameters: the initial alpha-wealth and the amount earned (called the pay-out) when a null hypothesis is rejected. Let $W(k) \geq 0$ denote the alpha-wealth accumulated by an investing rule after k tests; $W(0)$ is the initial alpha-wealth. For example, one might conventionally set $W(0) = 0.05$ or 0.10. At step j, an alpha-investing rule sets the level for testing H_j to some value α_j up to its current wealth, $0 \leq \alpha_j \leq W(j-1)$. The level α_j for testing H_j typically depends upon the sequence of prior outcomes R_1, R_2, $\ldots, R_{j-1}$, and so we write an alpha-investing rule in general as

$$\begin{aligned} \alpha_j &= \mathcal{I}_{W(0),\omega}(R_1,\, R_2, \ldots, R_{j-1}) \\ &= \mathcal{I}_{W(0),\omega}(j) \; . \end{aligned} \tag{7}$$

The outcomes of the sequence of tests determine the alpha-wealth $W(j-1)$ available for testing H_{j+1}. Let p_j denote the p-value of the test of H_j. If $p_j \leq \alpha_j$, the test rejects H_j. In this case, the investing rule pays $\log 1/(1-p_j) \approx p_j$ from the invested α_j and earns a pay-out ω that is added to its alpha-wealth. If $p_j > \alpha_j$, the procedure does not reject H_j and its alpha-wealth decreases by $\log(1-\alpha_j)$. The change in the alpha-wealth is thus

$$W(j) - W(j-1) = \begin{cases} \omega + \log(1-p_j) & \text{if } p_j \leq \alpha_j \;, \\ \log(1-\alpha_j) & \text{if } p_j > \alpha_j \;. \end{cases} \tag{8}$$

The appearance of $\log(1-\alpha)$ and $\log(1-p)$ in (8) deserves some explanation.

Consider the following “micro-investment” approach to testing a single null hypothesis H_0. Set the initial wealth $W(0) = \alpha$ and assume that the test of H_0 returns p-value p_0. Rather than use one test at level α, a micro-investment approach uses a sequence of tests, each risking a small amount $\epsilon \ll \alpha$ of the total alpha-wealth. First test H_0 at level ϵ, rejecting H_0 if $p_0 \leq \epsilon$. If $p_0 > \epsilon$, the investing rule pays ϵ for the first test, and then tests H_0 conditionally on $p_0 > \epsilon$ at level ϵ. This second test rejects H_0 if $\epsilon < p_0 \leq 2\epsilon - \epsilon^2$. If this second test does not reject H_0, the investing rule again pays ϵ and retests H_0, now conditionally on $p_0 > 2\epsilon - \epsilon^2$. This process continues until the investing rule either spends all of its alpha-wealth or rejects H_0 on the kth attempt because

$$1 - (1-\epsilon)^{k-1} < p_0 \leq 1 - (1-\epsilon)^k \; .$$

If the procedure rejects H_0 after k tests, then the total of the micro-payments made is

$$k\,\epsilon = \frac{\log(1-p_0)}{\log(1-\epsilon)}\,\epsilon \to -\log(1-p_0) \text{ as } \epsilon \to 0 \;.$$

The increments to the wealth defined in equation (8) essentially treat each test as a sequence of such micro-level tests.

In the next section, we show that alpha-investing rules that accumulate alpha-wealth in this way control EDC. The initial alpha-wealth $W(0)$ controls the chance for rejecting the null model. Under the null model when no hypothesis is rejected, an investing rule performs like an alpha-spending rule with level $W(0)$ and so $\text{Size}(m) \le W(0)$. Results described in the next section permit one to make a correspondence between the parameters $W(0)$ and ω that characterize an alpha-investing rule and the parameters α and γ that identify EDC. In particular, to control EDC, it will be shown most natural to associate $W(0)$ with α and ω with γ.

Whereas $W(0)$ controls the probability of rejecting the null model, the pay-out ω controls how the testing procedure performs once it has rejected the null model. The notion of compensation for rejecting a hypothesis captured in (8) allows one to build context-dependent information into the testing procedure. Suppose that the substantive context suggests that the first few hypotheses are most likely to be those that are rejected and that false hypotheses come in clusters. In this setting, one might consider using an alpha-investing rule like the following. Assume that the last rejected hypothesis is H_{k^*}. If false hypotheses are clustered, an alpha-investing rule should invest most of its wealth $W(k^*)$ available after rejecting H_{k^*} in testing H_{k^*+1}. A rule that does this is

$$\mathcal{I}_{W(0),\omega}(k) = \frac{6\,W(k^*)}{\pi^2}\;\frac{1}{(k-k^*)^2}\;, \quad k = k^*+1,\ldots,\min\{j : j > k^*, R_j = 1\}\;. \tag{9}$$

This rule invests $6/\pi^2 \approx 0.6$ of its wealth in testing H_1 or the null hypothesis H_{k^*+1} that follows a rejected hypothesis. The α-level falls off rapidly at the rate $1/k^2$ as more subsequent hypotheses are tested and not rejected. If the substantive insight is correct and the false null hypotheses are clustered, then tests of hypotheses like H_1 or H_{k^*+1} represent “good investments.” An example in Section 6 illustrates these ideas.

While it is relatively straightforward to devise investing rules, it may be difficult *a priori* to order the hypotheses in such a way that those most likely to be rejected come first. Such an ordering relies heavily on the structure of the specific testing situation. Another complication is the construction of tests that provide the p-values that determine the alpha-wealth of an investing rule according to (8). In order to show that a procedure controls EDC, we require a test of H_j to have the property that

$$\forall \theta \in \Theta, \quad E_\theta(V_j^\theta \mid R_{j-1}, R_{j-2}, \ldots, R_1) \leq \alpha_j \ . \tag{10}$$

This condition amounts to requiring that, conditionally on having either accepted or rejected the prior $j-1$ hypotheses, the test of H_j is done at level no higher than the nominal choice α_j. The tests need not be independent.

Remark. These procedures only require that the test of H_j maintain the stated α-level conditionally on the binary random variables $R_1, R_2, \ldots, R_{j-1}$. In particular, we note that the test is not conditioned on the test statistic (such as a z-score) or parameter estimate. Adaptive testing in a group sequential trial (e.g. Lehmacher and Wassmer, 1999) uses the information on the observed z-statistic at the first look. Tsiatis and Mehta (2003) shows that using this information leads to a less powerful test compared to traditional group sequential tests that only look at acceptance at the first look.

5 Alpha-Investing Rules Control EDC

An important extension of EDC generalizes this criterion to an arbitrary number of hypotheses. This version of the criterion replaces the fixed count of hypotheses in the definition (5) of $\mathrm{EDC}_{\alpha,\gamma}(m)$ by an arbitrary stopping time.

Definition 2. The excess discovery count of a procedure for testing a stream of hypotheses $H_1, H_2, \ldots$ is

$$\mathrm{EDC}_{\alpha,\gamma} = \inf_{\theta \in \Theta} \inf_{M \in \mathcal{M}} E_\theta \left(S^\theta(M) - \gamma R(M) \right) + \alpha \ . \tag{11}$$

where $M \in \mathcal{M}$, the set of stopping times with finite expectation.

The condition on $\mathcal{M}$ forces $S^\theta(M) \leq R(M) \leq M$ and so implies that both $E_\theta\, R(M)$ and $E_\theta\, S^\theta(M)$ are bounded. Because step-up testing halts after the last significant test (which is not a stopping time), this extension of EDC does not apply to such procedures. In what follows, we will concentrate then on step-down procedures.

We offer two observations on this generalized criterion. First, EDC drifts to $-\infty$ as the number of tests increases for any testing procedure that fixes the level of significance. To see that this is so, suppose a sequence of tests are made at level α (as in the naive procedure considered in the prior example). Under the null model, we expect

$100\alpha\%$ of the hypotheses to be falsely rejected. Because all of the null hypotheses are true, $S^\theta(m) = 0$ and $\mathrm{EDC}_{\alpha,\gamma}(m) = \alpha - \gamma\, E_\theta R(m) = \alpha(1 - \gamma m) \to -\infty$ as $m \to \infty$. Hence $\mathrm{EDC}_{\alpha,\gamma} = -\infty$.

Second, we observe that it is always possible to construct a test procedure for which $\mathrm{EDC}_{\alpha,\gamma} \geq 0$. The Bonferroni procedure offers a concrete example. Although the common application of the Bonferroni rule assigns equal α-level to each test, this need not be the case. All that is necessary is that the sum of the levels be less than α. If one tests H_j at level α_j and $\sum_j \alpha_j \leq \alpha$, then $E_\theta\ V^\theta(m) \leq \alpha$ for all m. Thus, $EDC_{\alpha,\gamma}(m) \geq 0$ for all α and m.

The following theorem states that an alpha-investing rule $\mathcal{I}_{W(0),\omega}$ with wealth determined by (8) controls EDC so long as the pay-out ω is not too large. The theorem follows by showing that a stochastic process related to the alpha-wealth sequence $W(0), W(1), \ldots$ is a sub-martingale. Because the proof of this result relies only on the optional stopping theorem for martingales, we do not require independent tests, though this is the certainly the easiest context in which to show that the p-values are honest in the sense required for (10) to hold.

Theorem 1 *An alpha-investing rule $\mathcal{I}_{W(0),\omega}$ governed by (8) with initial alpha-wealth $W(0) \leq \alpha$ and pay-out $\omega \leq 1 - \gamma$ controls $EDC_{\alpha,\gamma}$,*

$$EDC_{\alpha,\gamma} \geq 0\ . \tag{12}$$

A proof of the theorem is in the appendix.

6 Examples

The examples in this section illustrate alpha-investing rules and EDC. Our first two examples consider testing a large, but fixed, collection of m hypotheses for which we observe independent p-values p_1, p_2, $\ldots$, p_m. The first describes an alpha-investing rule that mimics Simes-based step-down testing. The second shows how alpha-investing rules are able to leverage domain knowledge to form a more powerful multiple testing procedure. A third example describes alpha-investing when testing a stream of hypothesis using dependent test statistics.

6.1 Comparison to Step-Down Testing

We compare alpha-investing to the Simes-based step-down testing procedure described in Section 2. This procedure rejects $H_{(1)}, H_{(2)}, \ldots, H_{(j_d^*-1)}$, where $j_d^* = \min\{k : p_{(k)} > k\,\alpha/m\}$ identifies the first test that is not rejected. (Step-up testing does not provide a stopping time.) Assume that the step-down procedure controls $\mathrm{FDR}(m) \le \alpha$ and rejects a small number $k > 0$ of the m hypotheses. It follows then that the p-values have the following structure:

$$p_{(1)} \le \alpha/m,\ p_{(2)} \le 2\alpha/m,\ \ldots, p_{(k)} \le k\alpha/m,\ \text{and}\ p_{(k+1)} > (k+1)\alpha/m\ . \qquad (13)$$

To reproduce this behavior with alpha-investing, consider the following approach. Set the initial alpha-wealth $W(0) = \alpha$ and $\omega = \alpha$. Define the alpha-investing to allocate its available alpha-wealth $W(j)$ equally over the hypotheses that have not been rejected, and begin by testing each hypothesis at the Bonferroni level α/m. Because of the structure in the p-values (13), this first pass rejects at least one hypothesis, namely $H_{(1)}$. To keep the presentation simple, suppose that only one hypothesis has p-value less than α/m. The procedure pays $\log(1-\alpha/m)$ for each test that does not reject, and earns $\alpha + \log(1-p_{(1)})$ for rejecting $H_{(1)}$. Hence, after testing each hypothesis at level α/m, its alpha-wealth is at least

$$\begin{aligned} W(m) &= W(0) + \alpha + \log(1-p_{(1)}) + (m-1)\log(1-\alpha/m) \\ &\ge 2\alpha + m\log(1-\alpha/m) \\ &\ge \alpha - \alpha^2/m \qquad (14) \end{aligned}$$

After this first pass through the hypotheses, its alpha-wealth is virtually unchanged, and it retains enough wealth to reject $H_{(2)}$.

For the second pass through the remaining $m-1$ null hypotheses, the alpha-investing rule rejects any hypothesis for which $p_j \le 2\alpha/m$, as in the Simes procedure. Because these tests condition on $p_j > \alpha/m$, this round of testing requires that the alpha-investing rule test each of the remaining $m-1$ hypothesis at level

$$\mathrm{P}_0\left(\frac{\alpha}{m} < p_j \le \frac{2\,\alpha}{m} \mid p_j > \frac{\alpha}{m}\right) = \frac{\alpha}{m-\alpha}\ .$$

It possesses enough wealth after the second round to do this because, from (14) for $\alpha \le 1/2$,

$$\frac{W(m)}{m-1} \ge \frac{\alpha - \alpha^2/m}{m-1} \ge \frac{\alpha}{m-\alpha}\ .$$

As in the first round, this second pass again approximately conserves the alpha-wealth of the procedure. Thus, so long as m is large and $k \ll m$ so that bounds similar to (14) hold, each pass though the hypotheses conserves enough alpha-wealth for the next round of tests. In this way, the investing rule gradually raises the threshold for rejecting a hypothesis as the number of rejected hypotheses increases.

The simulation summarized in the next section compares this alpha-investing rule to step-down testing. The alpha-investing rule generally does slightly better (rejects more false hypotheses) than step-down testing for two reasons. First, the lower bound (14) for the wealth $W(m)$, for example, assumes $p_{(1)} = \alpha/m$. In fact, we would expect $p_{(1)}$ to be closer to $\alpha/(2m)$, on average. Second, our description assumes that the p-values rejected by step-down testing are evenly distributed, with one between each threshold. Instead, it is likely that some passes of the investing rule will reject more than one hypothesis and thus have greater alpha-wealth for testing in the next round than suggested by these lower bounds.

6.2 Investing Rules that Leverage Domain Knowledge

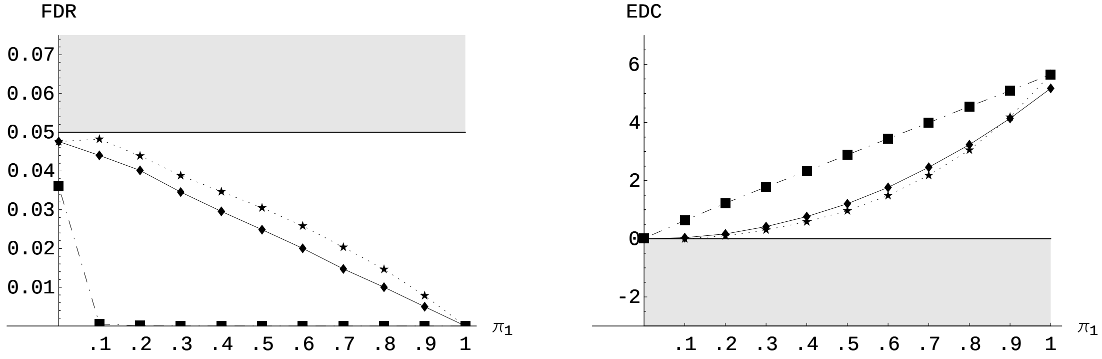
The performance of an alpha-investing rule improves, in the sense of being more powerful, if the investigator "knows the science". If the investigator is able to order the hypotheses *a priori* so that those most likely to be rejected are tested first, then alpha-investing can reject considerably more hypotheses than step-down testing. The full benefit is only realized, however, when one exploits an aggressive investing rule. The prior investing rule assumes that the hypotheses are arranged in no particular order and spreads its alpha-wealth evenly over the remaining hypotheses.

Suppose that the test procedure rejects H_{k^*} and is about to test H_{k^*+1}. Rather than spread its current alpha-wealth $W(k^*)$ evenly over the remaining hypotheses, a rule can invest more in testing the next hypothesis. For example, one can allocate $W(k^*)$ using a discrete probability mass function such as this version of the investing rule (9). If none of the remaining hypotheses are rejected, then the level for testing H_j is

$$\alpha_j = \frac{W(k^*)}{h_{m-k^*,2}} \, \frac{1}{(j-k^*)^2}, \quad j = k^*+1, \ldots, m \,, \tag{15}$$

where the normalizing constant $h_{q,2} = \sum_{i=1}^{q} 1/i^2$. If one of these tests rejects a hypothesis, the procedure reallocates its wealth so that all is spent by the time the procedure tests H_m. Mimicing the language of financial investing, we describe this type of alpha-

Figure 4: *Both alpha-investing rules (conservative —, aggressive $- \cdot -$) control FDR (left) and EDC (right), as does step-down testing ($\cdots$). Conservative alpha-investing assumes no domain knowledge, whereas aggressive alpha-investing uses domain knowledge, here the ordering of μ_i^2.*

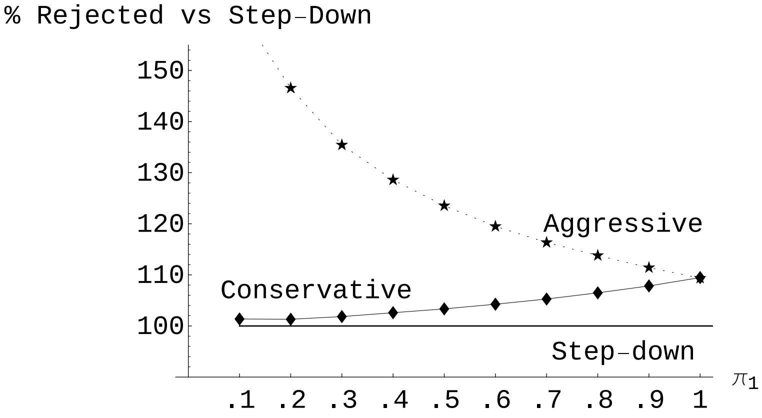


investing rule as aggressive and the previous method as conservative.

The simulation summarized in Figure 4 compares step-down testing to conservative and aggressive alpha-investing rules. For this simulation, we assume that the investigator tests the hypotheses in the order implied by $|\mu_j|$. The $m = 200$ hypotheses test means as defined in the simulation in Section 3 (see equation 6). We set the initial wealth $W(0) = 0.05$, $\alpha = 0.05$, $\gamma = 0.95$, and used step-down testing that controls $FDR(200) < 0.05$. Figure 4 shows FDR and EDC. All three procedures control both FDR and EDC, as they should. FDR for step-down testing closely tracks the performance of the conservative alpha-investing rule. The particularly low FDR obtained by aggressive alpha-investing may appear surprising at first. The low error rate is another benefit of the side-information. Aggressive alpha-investing spends all of its alpha-wealth testing the initial hypotheses — which happen to be false — and runs out of wealth before encountering the hypotheses for which $\mu_j = 0$. This rule also has larger EDC.

Alpha-investing guarantees protection from too many false rejections, but how well does it find signal? Figure 5 compares the power of the these alpha-investing rules to that of step-down testing. The plot shows the number of correct rejections $S^\theta(m)$ made by three different rules: aggressive alpha-investing that exploits domain knowledge using the rule (15), conservative alpha-investing (which assumes a random order) and step-down testing. The figure shows the average number of hypotheses rejected by each investing rule relative to the number rejected by step-down testing, on a percentage

Figure 5: *Aggressive alpha-investing using (15) exploits domain knowledge to achieve higher power than Simes-based step-down testing. This plot shows the percentage of correctly rejected null hypotheses for each procedure, relative to step-down testing. Both alpha-investing rules have more power than step-down testing with the same size.*



scale. For example, with a weak signal ($\pi_1 = 0.10$),

$$100\,\frac{S^\theta(m, \text{aggressive investing})}{S^\theta(m, \text{step-down})} > 150\%$$

In general, for weak signals, aggressive alpha-investing identifies about 30% more false hypotheses than step-down testing. The two become more similar as signal strength grows (in the form of more false null hypotheses). As discussed in the prior section, conservative alpha-investing rejects a few more hypothesis, about 5-10%, than Simes-based step-down testing.

6.3 Dependent Tests

The previous examples illustrate EDC and alpha-investing rules when testing a closed set of m hypotheses using independent tests. For dependent tests, however, step-down testing does not guarantee control of FDR. In comparison, one can find alpha-investing rules that control EDC.

EDC itself makes no assumption of independence of the the tests, but does require that the tests be conditionally correct in the sense of (10). When hypothesis tests are independent, it is simple to assure that each test indeed has level α_j. One need only form each test as though only one hypothesis were being tested; the outcomes of the prior tests R_1, R_2 $\ldots, R_{j-1}$ do not affect its level. This condition is much more difficult to establish when the tests are dependent. Although EDC allows any sort of dependence, it may not be possible to construct tests that satisfy this condition without making assumptions on the form of the dependence.

In some cases, however, known properties of multivariate distributions suggest a suitable test procedure. For example, suppose that the test statistics $Y = (Y_1, \ldots, Y_m)$ for $\mathcal{H}(m)$ have a multivariate normal distribution with mean vector $\vec{\mu}$ and covariance matrix Σ, $Y \sim N(\vec{\mu}, \Sigma)$. In this case, Dykstra (1980) shows that

$$\mathrm{P}(|Y_m| < c_m \mid |Y_1| \le c_1, \ldots, |Y_{m-1}| \le c_{m-1}) \ge \mathrm{P}(|Y_m| \le c_m)\ . \tag{16}$$

Thus, so long as no prior two-sided hypothesis has been rejected, an α-level test of H_m that ignores the prior outcomes — as though they were independent — has level at least α. The procedure is conservative. If, however, some prior test rejects a null hypothesis, these results no longer hold.

In this case, the simplest way to ensure the level of a test is to remove the effect of the rejected hypothesis. If H_k, say, has been rejected, then one can guarantee (10) holds by constructing subsequent tests to be independent of Y_k *and* any $Y_j, j < k$ which is correlated with Y_k. By removing the information from the rejected test, the acceptance region for subsequent two-sided tests is a symmetric convex set around the origin and inequalities such as (16) hold.

For example, consider a balanced two-way analysis of variance with r row effects $\beta_{r,i}$ and c column effects $\beta_{c,j}$ with $\sum_i \beta_{r,i} = \sum_j \beta_{c,j} = 0$. Write the vector of row effects as $\vec{\beta}_r$ and the vector of column effects $\vec{\beta}_c$. For each cell of the design, we have n independent normally distributed observations Y_{ijk}

$$Y_{ijk} = \mu_0 + \beta_{r,i} + \beta_{c,j} + Z_{ijk}, \quad Z_{ijk} \overset{\text{iid}}{\sim} N(0, \sigma^2), k = 1, \ldots, n,$$

with known variance σ^2. Assume that the hypotheses to be tested have the form $H_j : \vec{\lambda}'_{r,j}\vec{\beta}_r = 0, \vec{\lambda}'_{c,j}\vec{\beta}_c = 0$. Standard results from linear models show that the usual tests of H_j and H_k are independent if $\vec{\lambda}'_{r,j}\vec{\lambda}_{r,k} = 0$ and $\vec{\lambda}'_{c,j}\vec{\lambda}_{c,k} = 0$. Suppose one begins with tests of the row effects ($\lambda_c = 0$). There are no constraints on the tests until rejecting a hypothesis, H_k say. At this point, one can commence testing column effects, ignoring the prior results for the row effects because these are orthogonal. One can continue testing other hypotheses among the row effects so long as $\vec{\lambda}_{r,j}$ is orthogonal to $\vec{\lambda}_{r,k}$.

A similar procedure can be used in stepwise regression. Consider the familiar forward stepwise search, seeking predictors of the response Y among X_1, X_2, $\ldots, X_m$

in a linear model

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \cdots + \beta_m X_{m,i} + Z_i, \quad Z_i \overset{\text{iid}}{\sim} N(0, \sigma^2) \ .$$

Assume that all of the variables have mean zero and $\beta_0 = 0$. Under the normal linear model with known error variance, (16) implies that tests of $H_j : \beta_j = 0$ based on the familiar z-scores for the predictors $Z_j = (X_j'Y)/(X_j'X_j)$ satisfy (10) until some H_k is rejected. For further tests, one can assure that (10) holds by sweeping X_k *and* all predictors among X_1, X_2, $\ldots, X_{k-1}$ that are correlated with X_k from the remaining predictors. In practice, most predictors are correlated with each other to some extent and this condition requires sweeping X_1, X_2, $\ldots, X_k$ from subsequent predictors. If we collect these k predictors into an $n \times k$ matrix X, then the subsequent predictors would be $\tilde{X}_j = (I - X'(X'X)^{-1}X)X_j$, $j = k+1, \ldots$. The resulting loss of variation in predictors suggests it would be prudent to at least partially "orthogonalize" the predictors prior to using this type of search.

7 Discussion

The combination of EDC with alpha-investing rules invites the use of adaptive strategies for testing multiple hypotheses. Rather than posit a fixed set of hypotheses in advance of analysis, one can offer a strategy for determining which hypotheses to test next after getting some preliminary results. We would expect good strategies to leverage domain knowledge and be specific to the particular method of analysis.

Part of our motivation for developing EDC and alpha-investing rules arose from our work using stepwise regression for data mining (Foster and Stine, 2004). In this application, we compared forward stepwise regression to tree-based classifiers for predicting the onset of personal bankruptcy. To make regression competitive, we expanded the stepwise search to include all possible interactions among more than 350 "base" predictors. This produced more than 67,000 possible predictors. Because so many of these predictors were interactions (more than 98%), it is not surprising that most of the predictors identified by the search were interactions. Furthermore, because of the wide scope of this search, the procedure lacked power to find subtle effects that while small, improve the predictive power of a model. It became apparent to us that a hybrid search that only considered the interaction $X_j * X_k$, say, *after* including both X_j and X_k as main effects might be very effective. At the time, however, we lacked a method for controlling the selection procedure when the scope of the search dynamically expands as in this situation. We expect to use alpha-investing heavily in this work in the future.

We speculate that the greatest reward from developing a specialized testing strategy will come from developing methods that select the next hypothesis rather than specific functions to determine how α is spent. The rule (15) invests most of the current wealth in testing hypotheses following a rejection. One can imagine quite a few other choices. Our work and those of others in information theory (Rissanen, 1983; Foster, Stine and Wyner, 2002), however, suggest that one can find universal alpha-investing rules. Given a procedure for ordering the hypothesis, a universal alpha-investing rule would lead to rejecting as many hypothesis as the best rule within some class. We would expect such a rule to spend its alpha-wealth a bit more slowly than the simple rule (15), but retain this general form.

Another area of application for alpha-investing is in group-sequential clinical trials. In other work (Foster and Stine, 2005) we address the concept of adaptive design with a modification for alpha-investing. We show that the complaints raised in Tsiatis and Mehta (2003) about the efficiency of such tests can be mitigated by proper alpha-investing. At the same time, we allow the researcher freedom to design rules that guide how to spend or invest their alpha-wealth.

Appendix: Proof of Theorem 1

We prove Theorem 1 in this section. We begin by defining an empirical excess discovery count. Define the random variable

$$\mathrm{edc}_{\alpha,\gamma}(\theta, j) \equiv S^{\theta}(j) - \gamma R(j) + \alpha$$

so that

$$\mathrm{EDC}_{\alpha,\gamma} = \inf_{\theta\in\Theta} \inf_{M\in\mathcal{M}} E_{\theta}(\mathrm{edc}_{\alpha,\gamma}(\theta, M)) \ .$$

Now define

$$A(j) \equiv \mathrm{edc}_{\alpha,\gamma}(\theta, j) - W(j) \ .$$

Our main lemma shows that $A(j)$ is a sub-martingale for alpha-investing rules with initial alpha-wealth $W(0) \le \alpha$ and pay-out $\omega \le 1-\gamma$. A sub-martingale is “increasing” in the sense that

$$E_{\theta}\left(A(j) \mid A(j-1),\, A(j-2), \ldots, A(1)\right) \ge A(j-1) \ .$$

By definition $S^\theta(0) = R(0) = 0$ so that $\text{edc}_{\alpha,\gamma}(\theta, 0) = \alpha$. So if $W(0) \leq \alpha$, $\omega \leq 1-\gamma$ and $A(j)$ is a sub-martingale, then the optional stopping theorem implies that for all finite stopping times M

$$E_\theta(\text{edc}_{\alpha,\gamma}(\theta, M)) \geq E_\theta(\text{edc}_{\alpha,\gamma}(\theta, M) - W(M)) \geq \alpha - W(0) \geq 0 \ .$$

The first inequality follows because the alpha-wealth $W(j) \geq 0$ [*a.s.*], and the second inequality follows from the sub-martingale property. Since EDC for alpha-investing rules is the infimum over such expectations, all of which are non-negative, EDC itself is non-negative.

Thus to show Theorem 1 all we need is the following lemma:

Lemma 1 *Let $V^\theta(m)$ and $R(m)$ denote the cumulative number of false rejections and the cumulative number of all rejections, respectively, when testing a sequence of null hypotheses $\{H_1, H_2, \ldots\}$ using an alpha-investing rule $\mathcal{I}_{W(0),\omega}$ with initial alpha-wealth $W(0) \leq \alpha$, pay-out $\omega \leq 1-\gamma$, and cumulative alpha-wealth $W(m)$. Then the process*

$$\begin{aligned} A(j) &\equiv edc_{\alpha,\gamma}(\theta, j) - W(j) \\ &= (1-\gamma)R(j) - V^\theta(j) + \alpha - W(j) \end{aligned}$$

is a sub-martingale,

$$E\left(A(m) \mid A(m-1), \ldots, A(1)\right) \geq A(m-1) \ . \tag{17}$$

Proof.

We begin with some notation for the increments that define the counts in Table 1. Write $V^\theta(m)$ and $R(m)$ as sums of indicators V_j^θ, $R_j \in \{0, 1\}$,

$$V^\theta(m) = \sum_{j=1}^m V_j^\theta \ , \qquad R(m) = \sum_{j=1}^m R_j \ .$$

Similarly write the accumulated alpha-wealth $W(m)$ and $A(m)$ as sums of increments, $W(m) = \sum_{j=0}^m W_j$ and $A(m) = \sum_{j=0}^m A_j$. Let α_j denote the alpha level of the test of H_j that satisfies the condition (10). The change in the alpha-wealth from testing H_j can be written as:

$$W_j = R_j\omega + \log(1 - (p_j \wedge \alpha_j)) \ ,$$

where $\wedge$ is the minimum operator. Substituting this into A_j we get

$$A_j = (1 - \gamma - \omega)R_j - V_j^\theta - \log(1 - (p_j \wedge \alpha_j)) \ .$$

Since $R_j \geq 0$ and $1 - \gamma - \omega \geq 0$ by the conditions of the lemma, it follows that

$$A_j \geq -V_j^\theta - \log(1 - (p_j \wedge \alpha_j)) \ . \tag{18}$$

If $\theta_j \notin H_j$, then $V_j^\theta = 0$ and $A_j \geq 0$ almost surely. So we only need to consider the case in which the null hypothesis H_j is true.

Abbreviate the conditional expectation

$$E_\theta^{j-1}(X) = E_\theta \left(X \mid A(1),\, A(2),\, \ldots, A(j-1) \right) \ .$$

Then, when H_j is true, $p_j \sim U[0,1]$ so that

$$\begin{aligned} E_\theta^{j-1}(-\log(1-(p_j \wedge \alpha_j)) &= -\int_0^1 \log(1-(p \wedge \alpha_j))dp \\ &= -\int_0^{\alpha_j} \log(1-p)dp - \int_{\alpha_j}^1 \log(1-\alpha_j)dp \\ &= \alpha_j \ . \end{aligned}$$

Since $E_\theta^{j-1}(V_j^\theta) \leq \alpha_j$ by the definition of this being an α_j level test, equation (18) implies $E_\theta^{j-1}\, A_j \geq 0$.

□

References

Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statist. Soc., Ser. B*, **57**, 289–300.

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

Braun, H. I. (ed.) (1994) *The Collected Works of John W. Tukey: Multiple Comparisons*, vol. VIII. New York: Chapman & Hall.

Dudoit, S., Shaffer, J. P. and Boldrick, J. C. (2003) Multiple hypothesis testing in microarray experiments. *Statistical Science*, **18**, 71–103.

Dykstra, R. L. (1980) Product inequalities involving the multivariate normal-distribution. *Journal of the Amer. Statist. Assoc.*, **75**, 646–650.

Efron, B. (2005a) Large scale simultaneous hypothesis testing: the choice of a null hypothesis. *Journal of the Amer. Statist. Assoc.*, **100**, 96–104.

— (2005b) Selection and estimation for large-scale simultaneous inference. *Tech. rep.*, Department of Statistics, Stanford University, http://www-stat.stanford.edu/brad/papers/hivdata.

Foster, D. P. and Stine, R. A. (2004) Variable selection in data mining: Building a predictive model for bankruptcy. *Journal of the Amer. Statist. Assoc.*, **99**, 303–313.

— (2005) Theoretical foundations for adaptive testing using alpha-investing rules. *Tech. rep.*, Statistics Department, University of Pennsylvania.

Foster, D. P., Stine, R. A. and Wyner, A. J. (2002) Universal codes for finite sequences of integers drawn from a monotone distribution. *IEEE Trans. on Info. Theory*, **48**, 1713–1720.

Gupta, M. and Ibrahim, J. G. (2005) Towards a complete picture of gene regulation: using Bayesian approaches to integrate genomic sequence and expression data. *Tech. rep.*, University of North Carolina, Chapel Hill, NC.

Holm, S. (1979) A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, **6**, 65–70.

Lehmacher, W. and Wassmer, G. (1999) Adaptive sample size calculations in group sequential trials. *Biometrics*, **55**, 1286–90.

Marcus, R., Peritz, E. and Gabriel, K. R. (1976) On closed testing procedures with special reference to ordered analysis of variance. *Biometrika*, **63**, 655–660.

Meinshausen, N. and Buehlmann, P. (2004) Lower bounds for the number of false null hypotheses for multiple testing of associations under general dependence. *Tech. Rep. 121*, ETH Zurich, http://stat.ethz.ch/ nicolai/.

Meinshausen, N. and Rice, J. (2004) Estimating the proportion of false null hypotheses among a large number of independently tested hypotheses. To appear, *Annals of Statistics*.

Rissanen, J. (1983) A universal prior for integers and estimation by minimum description length. *Annals of Statistics*, **11**, 416–431.

Sarkar, S. K. (1998) Some probability inequalities for ordered Mtp$_2$ random variables: A proof of the Simes conjecture. *Annals of Statistics*, **26**, 494–504.

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

Storey, J. D. (2002) A direct approach to false discovery rates. *Journal of the Royal Statist. Soc., Ser. B*, **64**, 479–498.

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

Troendle, J. F. (1996) A permutation step-up method of testing multiple outcomes. *Biometrics*, **52**, 846–859.

Tsiatis, A. A. and Mehta, C. (2003) On the inefficiency of the adaptive design for monitoring clinical trials. *Biometrika*, **90**, 367–378.

Tukey, J. W. (1953) The problem of multiple comparisons. Unpublished lecture notes.

— (1991) The philosophy of multiple comparisons. *Statistical Science*, **6**, 100–116.