Statistical Science
2008, Vol. 23, No. 1, 1–22
DOI: 10.1214/07-STS236

Microarrays, Empirical Bayes and the Two-Groups Model

Bradley Efron

Abstract. The classic frequentist theory of hypothesis testing developed by Neyman, Pearson and Fisher has a claim to being the twentieth century's most influential piece of applied mathematics. Something new is happening in the twenty-first century: high-throughput devices, such as microarrays, routinely require simultaneous hypothesis tests for thousands of individual cases, not at all what the classical theory had in mind. In these situations empirical Bayes information begins to force itself upon frequentists and Bayesians alike. The two-groups model is a simple Bayesian construction that facilitates empirical Bayes analysis. This article concerns the interplay of Bayesian and frequentist ideas in the two-groups setting, with particular attention focused on Benjamini and Hochberg's False Discovery Rate method. Topics include the choice and meaning of the null hypothesis in large-scale testing situations, power considerations, the limitations of permutation methods, significance testing for groups of cases (such as pathways in microarray studies), correlation effects, multiple confidence intervals and Bayesian competitors to the two-groups model.

Key words and phrases: Simultaneous tests, empirical null, false discovery rates.

1. INTRODUCTION

Simultaneous hypothesis testing was a lively research topic during my student days, exemplified by Rupert Miller's classic text "Simultaneous Statistical Inference" (1966, 1981). Attention focused on testing N null hypotheses at the same time, where N was typically less than half a dozen, though the requisite tables might go up to $N = 20$. Modern scientific technology, led by the microarray, has upped the ante in dramatic fashion: my examples here will have N's ranging from 200 to 10,000, while N = 500,000, from SNP analyses, is waiting in the wings. [The astrostatistical applications in Liang et al. (2004) envision $N = 10^{10}$ and more!]

Miller's text is relentlessly frequentist, reflecting a classic Neyman–Pearson testing framework, with the main goal being preservation of "α," overall test size, in the face of multiple inference. Most of the current microarray statistics literature shares this goal, and also its frequentist viewpoint, as described in the nice review article by Dudoit and Boldrick (2003).

Something changes, though, when N gets big: with thousands of parallel inference problems to consider simultaneously, Bayesian considerations begin to force themselves even upon dedicated frequentists. The "two-groups model" of the title is a particularly simple Bayesian framework for large-scale testing situations. This article explores the interplay of frequentist and Bayesian ideas in the two-groups setting, with particular attention paid to False Discovery Rates (Benjamini and Hochberg, 1995).

Figure 1 concerns four examples of large-scale simultaneous hypothesis testing. Each example consists of N individual cases, with each case represented by its own z-value "z_i," for $i = 1, 2, \ldots, N$. The z_i's are based on familiar constructions that, theoretically, should yield standard $N(0, 1)$ normal distributions un-

Bradley Efron is Professor, Department of Statistics, Stanford University, Stanford, California 94305, USA (e-mail: brad@stat.stanford.edu).

[1]Discussed in 10.1214/07-STS236B, 10.1214/07-STS236C, 10.1214/07-STS236D and 10.1214/07-STS236A; rejoinder at 10.1214/08-STS236REJ.

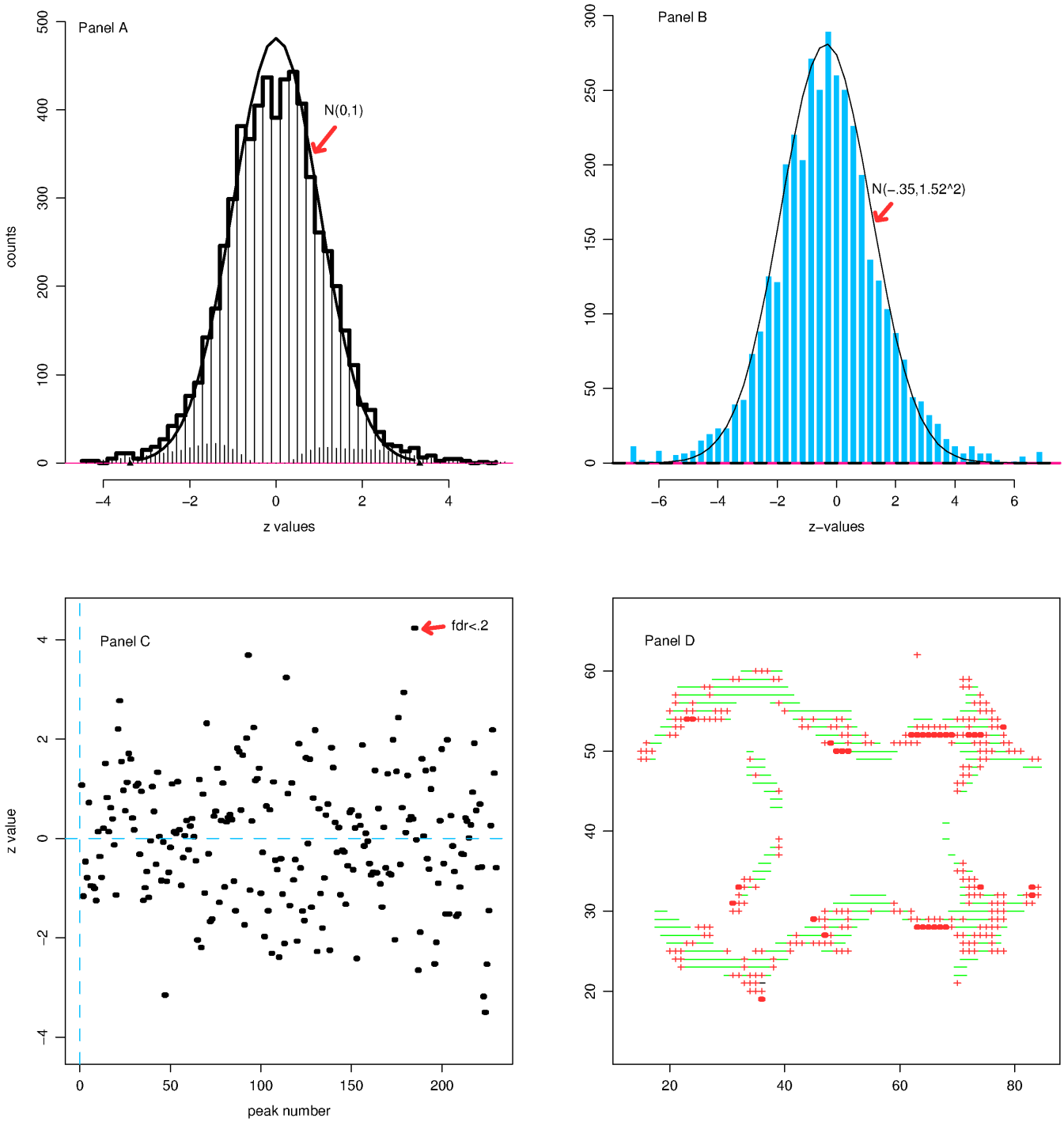

FIG. 1. *Four examples of large-scale simultaneous inference, each panel indicating N z-values as explained in the text. Panel* A, *prostate cancer microarray study, $N = 6033$ genes; panel* B, *comparison of advantaged versus disadvantaged students passing mathematics competency tests, $N = 3748$ high schools; panel* C, *proteomics study, $N = 230$ ordered peaks in time-of-flight spectroscopy experiment; panel* D, *imaging study comparing dyslexic versus normal children, showing horizontal slice of* 655 *voxels out of $N = 15{,}455$, coded "−" for $z_i < 0$, "+" for $z_i \geq 0$ and solid circle for $z_i > 2$.*

der a classical null hypothesis,

$$\text{theoretical null}: z_i \sim N(0,1). \tag{1.1}$$

Here is a brief description of the four examples, with further information following as needed in the sequel.

EXAMPLE A [Prostate data, Singh et al. (2002)]. $N = 6033$ genes on 102 microarrays, $n_1 = 50$ healthy males compared with $n_2 = 52$ prostate cancer patients; z_i's based on two-sample t statistics comparing the two categories.

EXAMPLE B [Education data, Rogosa (2003)]. $N = 3748$ California high schools; z_i's based on binomial test of proportion advantaged versus proportion disadvantaged students passing mathematics competency tests.

EXAMPLE C [Proteomics data, Turnbull (2006)]. $N = 230$ ordered peaks in time-of-flight spectroscopy study of 551 heart disease patients. Each peak's z-value was obtained from a Cox regression of the patients' survival times, with the predictor variable being the 551 observed intensities at that peak.

EXAMPLE D [Imaging data, Schwartzman et al. (2005)]. $N = 15{,}445$ voxels in a diffusion tensor imaging (DTI) study comparing 6 dyslexic with six normal children; z_i's based on two-sample t statistics comparing the two groups. The figure shows only a single horizontal brain section having 655 voxels, with "$-$" indicating $z_i < 0$, "$+$" for $z_i \geq 0$, and solid circles for $z_i > 2$.

Our four examples are enough alike to be usefully analyzed by the two-groups model of Section 2, but there are some striking differences, too: the theoretical $N(0, 1)$ null (1.1) is obviously inappropriate for the education data of panel B; there is a hint of correlation of z-value with peak number in panel C, especially near the right limit; and there is substantial spatial correlation appearing in the imaging data of panel D.

My plan here is to discuss a range of inference problems raised by large-scale hypothesis testing, many of which, it seems to me, have been more or less underemphasized in a literature focused on controlling Type-I errors: the choice of a null hypothesis, limitations of permutation methods, the meaning of "null" and "nonnull" in large-scale settings, questions of power, test of significance for groups of cases (e.g., pathways in microarray studies), the effects of correlation, multiple confidence statements and Bayesian competitors to the two-groups model. The presentation is intended to be as nontechnical as possible, many of the topics being discussed more carefully in Efron (2004, 2005, 2006). References will be provided as we go along, but this is not intended as a comprehensive review. Microarrays have stimulated a burst of creativity from the statistics community, and I apologize in advance for this article's concentration on my own point of view, which aims at minimizing the amount of statistical modeling required of the statistician. More model-intensive techniques, including fully Bayesian approaches, as in Parmigiani et al. (2002) or Lewin et al. (2006), have their own virtues, which I hope will emerge in the Discussion.

Section 2 discusses the two-groups model and false discovery rates in an idealized Bayesian setting. Empirical Bayes methods are needed to carry out these ideas in practice, as discussed in Section 3. This discussion assumes a "good" situation, like that of Example A, where the theoretical null (1.1) fits the data. When it does not, as in Example B, the *empirical null* methods of Section 4 come into play. These raise interpretive questions of their own, as mentioned above, discussed in the later sections.

We are living through a scientific revolution powered by the new generation of high-throughput observational devices. This is a wonderful opportunity for statisticians, to redemonstrate our value to the scientific world, but also to rethink basic topics in statistical theory. Hypothesis testing is the topic here, a subject that needs a fresh look in contexts like those of Figure 1.

2. THE TWO-GROUPS MODEL AND FALSE DISCOVERY RATES

The two-groups model is too simple to have a single identifiable author, but it plays an important role in the Bayesian microarray literature, as in Lee et al. (2000), Newton et al. (2001) and Efron et al. (2001). We suppose that the N cases ("genes" as they will be called now in deference to microarray studies, though they are not genes in the last three examples of Figure 1) are each either *null* or *nonnull* with prior probability p_0 or $p_1 = 1 - p_0$, and with z-values having density either $f_0(z)$ or $f_1(z)$,

$$\begin{aligned} p_0 &= \Pr\{\text{null}\} & f_0(z) &\text{ density if null,} \\ p_1 &= \Pr\{\text{nonnull}\} & f_1(z) &\text{ density if nonnull.} \end{aligned} \tag{2.1}$$

The usual purpose of large-scale simultaneous testing is to reduce a vast set of possibilities to a much smaller set of scientifically interesting prospects. In

Example A, for instance, the investigators were probably searching for a few genes, or a few hundred at most, worthy of intensive study for prostate cancer etiology. I will assume

$$p_0 \geq 0.90 \tag{2.2}$$

in what follows, limiting the nonnull genes to no more than 10%.

False discovery rate (Fdr) methods have developed in a strict frequentist framework, beginning with Benjamini and Hochberg's seminal 1995 paper, but they also have a convincing Bayesian rationale in terms of the two-groups model. Let $F_0(z)$ and $F_1(z)$ denote the cumulative distribution functions (cdf) of $f_0(z)$ and $f_1(z)$ in (2.1), and define the mixture cdf $F(z) = p_0F_0(z) + p_1F_1(z)$. Then Bayes' rule yields the a posteriori probability of a gene being in the null group of (2.1) given that its z-value Z is less than some threshold z, say "Fdr(z)," as

$$\begin{aligned}\mathrm{Fdr}(z) &\equiv \Pr\{\mathrm{null}|Z \leq z\} \\ &= p_0F_0(z)/F(z).\end{aligned} \tag{2.3}$$

[Here it is notationally convenient to consider the negative end of the z scale, values like $z = -3$. Definition (2.3) could just as well be changed to $Z > z$ or $Z > |z|$.] Benjamini and Hochberg's (1995) false discovery rate control rule begins by estimating $F(z)$ with the empirical cdf

$$\bar{F}(z) = \#\{z_i \leq z\}/N, \tag{2.4}$$

yielding $\overline{\mathrm{Fdr}}(z) = p_0F_0(z)/\bar{F}(z)$. The rule selects a control level "q," say $q = 0.1$, and then declares as nonnull those genes having z-values z_i satisfying $z_i \leq z_0$, where z_0 is the maximum value of z satisfying

$$\overline{\mathrm{Fdr}}(z_0) \leq q \tag{2.5}$$

[usually taking $p_0 = 1$ in (2.3), and F_0 the theoretical null, the standard normal cdf $\Phi(z)$ of (1.1)].

The striking theorem proved in the 1995 paper was that the expected proportion of null genes reported by a statistician following rule (2.5) will be no greater than q. This assumes independence among the z_i's, extended later to various dependence models in Benjamini and Yekutieli (2001). The theorem is a purely frequentist result, but as pointed out in Storey (2002) and Efron and Tibshirani (2002), it has a simple Bayesian interpretation via (2.3): rule (2.5) is essentially equivalent to declaring nonnull those genes whose estimated tail-area posterior probability of being null is no greater than q. It is usually a good sign when Bayesian and frequentist ideas converge on a single methodology, as they do here.

Densities are more natural than tail areas for Bayesian fdr interpretation. Defining the *mixture density* from (2.1),

$$f(z) = p_0f_0(z) + p_1f_1(z), \tag{2.6}$$

Bayes' rule gives

$$\begin{aligned}\mathrm{fdr}(z) &\equiv \Pr\{\mathrm{null}|Z = z\} \\ &= p_0f_0(z)/f(z)\end{aligned} \tag{2.7}$$

for the probability of a gene being in the null group given z-score z. Here fdr(z) is the *local false discovery rate* (Efron et al., 2001; Efron, 2005).

There is a simple relationship between Fdr(z) and fdr(z),

$$\mathrm{Fdr}(z) = E_f\{\mathrm{fdr}(Z)|Z \leq z\}, \tag{2.8}$$

"E_f" indicating expectation with respect to the mixture density $f(z)$. That is, Fdr(z) is the mixture average of fdr(Z) for $Z \leq z$. In the usual situation where fdr(z) decreases as $|z|$ gets large, Fdr(z) will be smaller than fdr(z). Intuitively, if we decide to label all genes with z_i less than some negative value z_0 as nonnull, then fdr(z_0), the false discovery rate at the boundary point z_0, will be greater than Fdr(z_0), the average false discovery rate beyond the boundary. Figure 2 illustrates the geometrical relationship between Fdr(z) and fdr(z); the Benjamini–Hochberg Fdr control rule amounts to an upper bound on the secant slope.

For Lehmann alternatives

$$F_1(z) = F_0(z)^\gamma, \quad [\gamma < 1], \tag{2.9}$$

it turns out that

$$\begin{aligned}&\log\left\{\frac{\mathrm{fdr}(z)}{1-\mathrm{fdr}(z)}\right\} \\ &\quad = \log\left\{\frac{\mathrm{Fdr}(z)}{1-\mathrm{Fdr}(z)}\right\} + \log\left(\frac{1}{\gamma}\right),\end{aligned} \tag{2.10}$$

so

$$\mathrm{fdr}(z) \doteq \mathrm{Fdr}(z)/\gamma \tag{2.11}$$

for small values of Fdr. The prostate data of Figure 1 has γ about $1/2$ in each tail, making fdr$(z) \sim 2$ Fdr(z) near the extremes.

The statistics literature has not reached consensus on the choice of q for the Benjamini–Hochberg control rule (2.5)—what would be the equivalent of 0.05 for classical testing—but Bayes factor calculations offer some insight. Efron (2005, 2006) uses the cutoff point

$$\mathrm{fdr}(z) \leq 0.20 \tag{2.12}$$

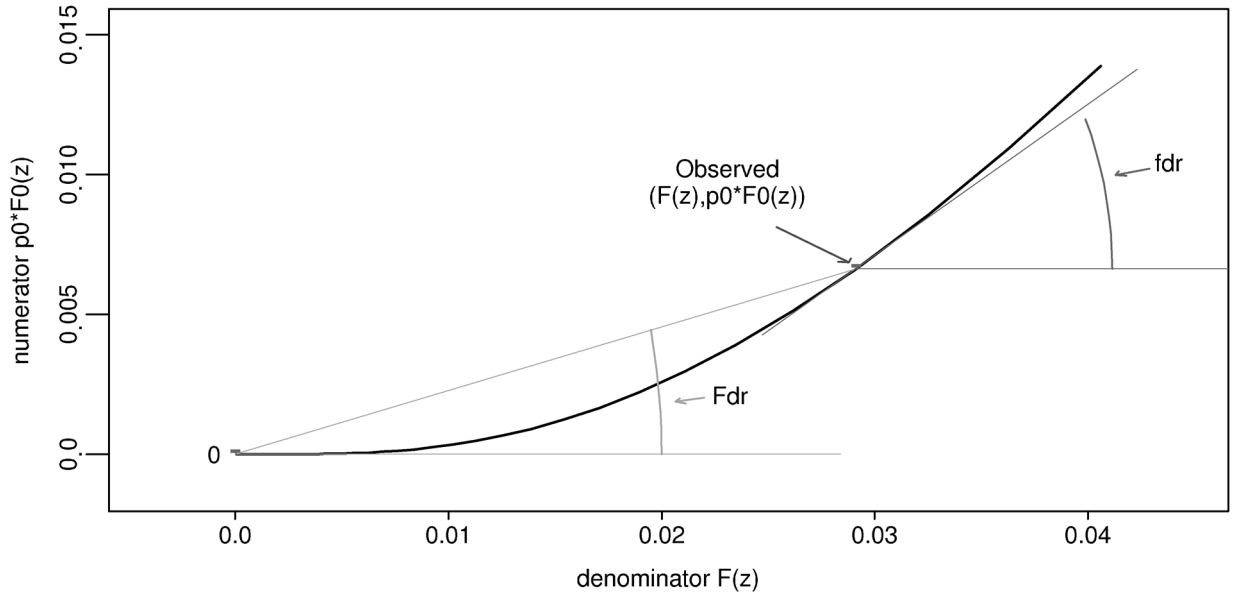

FIG. 2. *Relationship of Fdr(z) to fdr(z). Heavy curve plots numerator of Fdr, $p_0F_0(z)$, versus denominator $F(z)$; fdr(z) is slope of tangent, Fdr slope of secant.*

for reporting nonnull genes, on the admittedly subjective grounds that fdr values much greater than 0.20 are dangerously prone to wasting investigators' resources. Then (2.6), (2.7) yield posterior odds ratio

$$\begin{aligned}&\Pr\{\text{nonnull}|z\}/\Pr\{\text{null}|z\}\\&\quad=(1-\text{fdr}(z))/\text{fdr}(z)\\&\quad=p_1f_1(z)/p_0f_0(z)\\&\quad\geq 0.8/0.2=4.\end{aligned}\tag{2.13}$$

Since (2.2) implies $p_1/p_0 \leq 1/9$, (2.13) corresponds to requiring Bayes factor

$$f_1(z)/f_0(z) \geq 36 \tag{2.14}$$

in favor of nonnull in order to declare significance.

Factor (2.14) requires much stronger evidence against the null hypothesis than in standard one-at-a-time testing, where the critical threshold lies somewhere near 3 (Efron and Gous, 2001). The fdr 0.20 threshold corresponds to q-values in (2.5) between 0.05 and 0.15 for moderate choices of γ; such q-value thresholds can be interpreted as providing conservative Bayes factors for Fdr testing.

Model (2.1) ignores the fact that investigators usually begin with hot prospects in mind, genes that have high prior probability of being interesting. Suppose $p_0(i)$ is the prior probability that gene i is null, and define p_0 as the average of $p_0(i)$ over all N genes. Then Bayes' theorem yields this expression for $\text{fdr}_i(z) = \Pr\{\text{gene}_i \text{ null}|z_i = z\}$:

$$\begin{aligned}&\text{fdr}_i(z)=\text{fdr}(z)\frac{r_i}{1-(1-r_i)\text{fdr}(z)},\\&\qquad\left[r_i=\frac{p_0(i)}{1-p_0(i)}\Big/\frac{p_0}{1-p_0}\right],\end{aligned}\tag{2.15}$$

where $\text{fdr}(z) = p_0f_0(z)/f(z)$ as before. So for a hot prospect having $p_0(i) = 0.50$ rather than $p_0 = 0.90$, (2.15) changes an uninteresting result like $\text{fdr}(z_i) = 0.40$ into $\text{fdr}_i(z_i) = 0.069$.

Wonderfully neat and exact results like the Benjamini–Hochberg Fdr control rule exert a powerful influence on statistical theory, sometimes more than is good for applied work. Much of the microarray statistics literature seems to me to be overly concerned with exact properties borrowed from classical test theory, at the expense of ignoring the complications of large-scale testing. Neatness and exactness are mostly missing in what follows as I examine an empirical Bayes approach to the application of two-groups/Fdr ideas to situations like those in Figure 1.

3. EMPIRICAL BAYES METHODS

In practice, the difference between Bayesian and frequentist statisticians is their self-confidence in assigning prior distributions to complicated probability models. Large-scale testing problems certainly look

complicated enough, but this is deceptive; their massively parallel structure, with thousands of *similar* situations each providing information, allows an appropriate prior distribution to be estimated from the data without upsetting even timid frequentists like myself. This is the *empirical Bayes* approach of Robbins and Stein, 50 years old but coming into its own in the microarray era; see Efron (2003).

Consider estimating the local false discovery rate $\text{fdr}(z) = p_0 f_0(z)/f(z)$, (2.7). I will begin with a "good" case, like the prostate data of Example A in Section 1, where it is easy to believe in the theoretical null distribution (1.1),

$$f_0(z) = \varphi(z) \equiv \frac{1}{\sqrt{2\pi}} e^{-(1/2)z^2}. \tag{3.1}$$

The z-values in Example A were obtained by transforming the usual two-sample t statistic "t_i" comparing cancer and normal patients' expression levels for gene i, to a standard normal scale via

$$z_i = \Phi^{-1}(F_{100}(t_i)); \tag{3.2}$$

here Φ and F_{100} are the cdf's of standard normal and t_{100} distributions. If we had only gene i's data to test, classic theory would tell us to compare z_i with $f_0(z) = \varphi(z)$ as in (3.1).

For the moment I will take p_0, the prior probability of a gene being null, as known. Section 4 discusses p_0's estimation, but in fact its exact value does not make much difference to $\text{Fdr}(z)$ or $\text{fdr}(z)$, (2.3) or (2.7), if p_0 is near 1 as in (2.2). Benjamini and Hochberg (1995) take $p_0 = 1$, providing an upper bound for $\text{Fdr}(z)$.

This leaves us with only the denominator $f(z)$ to estimate in (2.7). By definition (2.6), $f(z)$ is the marginal density of all N z_i's, so we can use all the data to estimate $f(z)$. The algorithm *locfdr*, an *R* function available from the CRAN library, does this by means of standard Poisson GLM software (Efron, 2005). Suppose the z-values have been binned, giving bin counts

$$y_k = \#\{z_i \text{ in bin } k\}, \quad k = 1, 2, \ldots, K. \tag{3.3}$$

The prostate data histogram in panel A of Figure 1 has $K = 49$ bins of width $\Delta = 0.2$.

We take the y_k to be independent Poisson counts,

$$y_k \overset{\text{ind}}{\sim} P_0(\nu_k), \quad k = 1, 2, \ldots, K, \tag{3.4}$$

with the unknown ν_k proportional to density $f(z)$ at midpoint "x_k" of the kth bin, approximately

$$\nu_k = N\Delta f(x_k). \tag{3.5}$$

Modeling $\log(\nu_k)$ as a pth-degree polynomial function of x_k makes (3.4)–(3.5) a standard Poisson general linear model (GLM). The choice $p = 7$ used in Figure 3 amounts to estimating $f(z)$ by maximum likelihood within the seven-parameter exponential family

$$f(z) = \exp\left\{\sum_{j=0}^{7} \beta_j z^j\right\}. \tag{3.6}$$

Notice that $p = 2$ would make $f(z)$ normal; the extra parameters in (3.6) allow flexibility in fitting the tails of $f(z)$. Here we are employing *Lindsey's method*; see Efron and Tibshirani (1996). Despite its unorthodox look, it is no more than a convenient way to obtain maximum likelihood estimates in multiparameter families like (3.6).

The heavy curve in Figure 3 is an estimate of the local false discovery rate for the prostate data,

$$\widehat{\text{fdr}}(z) = p_0 f_0(z)/\widehat{f}(z), \tag{3.7}$$

with $\widehat{f}(z)$ constructed as above, $f_0(z) = \varphi(z)$ as in (3.1), and $p_0 = 0.93$, as estimated in Section 4; $\widehat{\text{fdr}}(z)$ is near 1 for $|z| \le 2$, decreasing to interesting levels for $|z| > 3$. Fifty-one of the 6033 genes have $\widehat{\text{fdr}}(z_i) \le 0.2$, 26 on the right and 25 on the left, and these could be reported back to the investigators as likely nonnull candidates. [The standard Benjamini–Hochberg procedure, (2.5) with $q = 0.1$, reports 60 nonnull genes, 28 on the right and 32 on the left.]

At this point the reader might notice an anomaly: if $p_0 = 0.93$ of the N genes are null, then about $(1 - p_0) \cdot 6033 = 422$ should be nonnull, but only 51 are reported. The trouble is that most of the nonnull genes are located in regions of the z axis where $\widehat{\text{fdr}}(z_i)$ exceeds 0.5, and these cannot be reported without also reporting a bevy of null cases. In other words, the prostate study is underpowered.

The vertical bars in Figure 3 are estimates of the *nonnull counts*, the histogram we would see if only the nonnull genes provided z-values. In terms of (3.3), (3.7), the nonnull counts "$y_k^{(1)}$" are

$$y_k^{(1)} = [1 - \widehat{\text{fdr}}_k]y_k, \tag{3.8}$$

where $\widehat{\text{fdr}}_k = \widehat{\text{fdr}}(x_k)$, the estimated fdr value at the center of bin k. Since $1 - \widehat{\text{fdr}}_k$ approximates the nonnull probability for a gene in bin k, formula (3.8) is an obvious estimate for the expected number of nonnulls.

Power diagnostics are obtained from comparisons of $\widehat{\text{fdr}}(z)$ with the nonnull histogram. High power would be indicated if $\widehat{\text{fdr}}_k$ was small where $y_k^{(1)}$ was large.

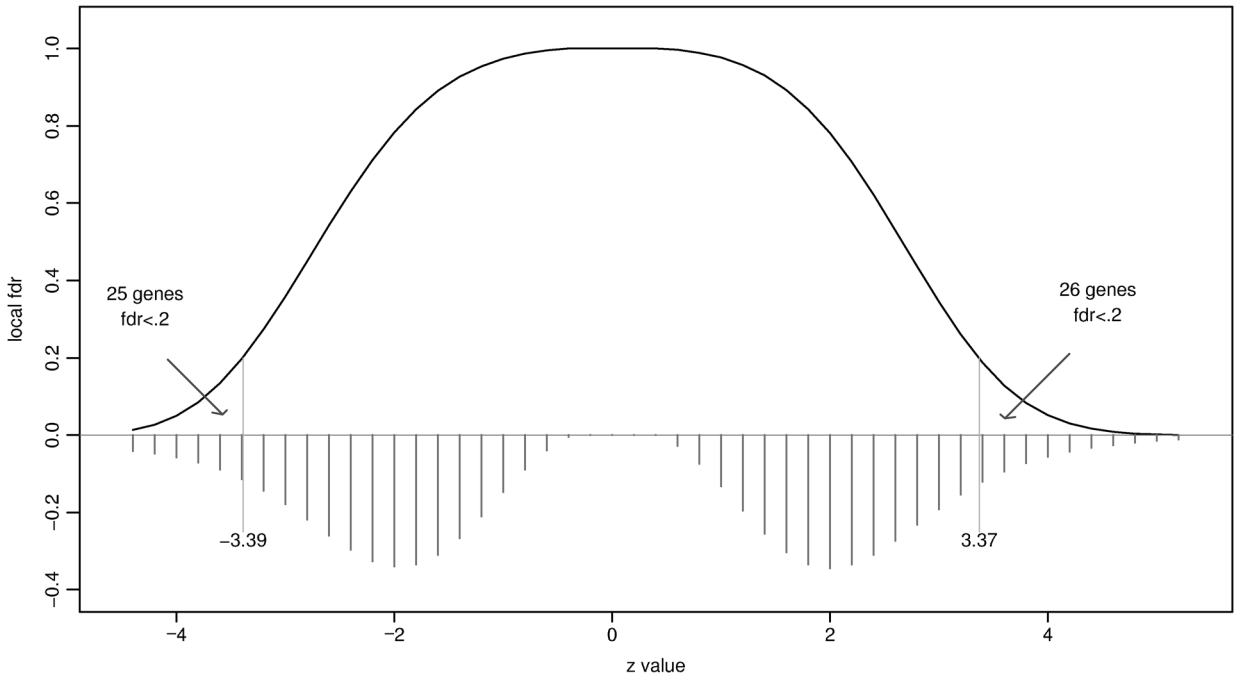

FIG. 3. *Heavy curve is estimated local false discovery rate* $\widehat{fdr}(z)$ *for prostate data. Fifty-one genes,* 26 *on the right and* 25 *on the left, have* $\widehat{fdr}(z_i) < 0.20$. *Vertical bars estimate histogram of the nonnull counts (plotted negatively, divided by* 50). *Most of the nonnull genes will not be reported.*

That obviously is not the case in Figure 3. A simple power diagnostic is

$$\widehat{E\ \mathrm{fdr}}^{(1)} = \sum_{k=1}^{K} y_k^{(1)} \widehat{\mathrm{fdr}}_k \Big/ \sum_{k=1}^{K} \widehat{y}_k^{(1)}, \tag{3.9}$$

the expected nonnull fdr. We want $\widehat{E\ \mathrm{fdr}}^{(1)}$ to be small, perhaps near 0.2, so that a typical nonnull gene will show up on a list of likely prospects. The prostate data has $\widehat{E\ \mathrm{fdr}}^{(1)} = 0.68$, indicating low power. If the whole study were rerun, we could expect a different list of 50 likely nonnull genes, barely overlapping with the first list. Section 3 of Efron (2006) discusses power calculations for microarray studies, presenting more elaborate power diagnostics.

Stripped of technicalities, the idea underlying false discovery rates is appealingly simple, and in fact does not depend on the literal validity of the two-groups model (2.1). Consider the bin $z_i \in [3.1, 3.3]$ in the prostate data histogram; 17 of the 6033 genes fall into this bin, compared to expected number $2.68 = p_0 N \Delta\varphi(3.2)$ of null genes, giving

$$\overline{\mathrm{fdr}} = 2.68/17 = 0.16 \tag{3.10}$$

as an estimated false discovery rate. (The smoothed estimate in Figure 3 is $\widehat{\mathrm{fdr}} = 0.24$.) The implication is that only about one-sixth of the 17 are null genes. This conclusion can be sharpened, as in Lehmann and Romano (2005), but (3.10) catches the main idea.

Notice that we do not need all the null genes to have the *same* density $f_0(z)$; it is enough to assume that the *average* null density is $f_0(z)$, $\varphi(z)$ in this case, in order to calculate the numerator 2.68. (This is an advantage of false discovery rate methods, which only control *rates*, not *individual probabilities*.) The nonnull density $f_1(z)$ in (2.1) plays no role at all since the denominator 17 is an observed quantity. *Exchangeability* is the key assumption in interpreting (3.10): we expect about $1/6$ of the 17 genes to be null, and assign posterior null probability $1/6$ to all 17. Nonexchangeability, in the form of differing prior information among the 17, can be incorporated as in (2.15).

Density estimation has a reputation for difficulty, well-deserved in general situations. However, there are good theoretical reasons, presented in Section 6 of Efron (2005), for believing that mixtures of z-values are quite smooth, and that (3.7) will efficiently estimate $\mathrm{fdr}(z)$. Independence of the z_i's is *not* required, only that $\widehat{f}(z)$ is a reasonably close estimate of $f(z)$.

TABLE 1
Boldface, standard errors of $\log \widehat{\text{fdr}}(z)$, *(local fdr), and* $\log \widehat{\text{Fdr}}(z)$, *(tail-area), 250 replications of model (3.11),* $N = 1500$. *Parentheses, average from formula (5.9), Efron (2006); fdr is true value (2.7). Empirical Null results explained in Section 4*

		Theoretical null			Empirical null		
z	fdr	local	(formula)	tail	local	(formula)	tail
1.5	0.88	**0.05**	(0.05)	**0.05**	**0.04**	(0.04)	**0.10**
2.0	0.69	**0.08**	(0.09)	**0.05**	**0.09**	(0.10)	**0.15**
2.5	0.38	**0.09**	(0.10)	**0.05**	**0.16**	(0.16)	**0.23**
3.0	0.12	**0.08**	(0.10)	**0.06**	**0.25**	(0.25)	**0.32**
3.5	0.03	**0.10**	(0.13)	**0.07**	**0.38**	(0.38)	**0.42**
4.0	0.005	**0.11**	(0.15)	**0.10**	**0.50**	(0.51)	**0.52**

Table 1 reports on a small simulation study in which

$$z_i \overset{\text{ind}}{\sim} N(\mu_i, 1) \begin{cases} \mu_i = 0, \\ \quad \text{with probability } 0.9, \\ \mu_i \sim N(3, 1), \\ \quad \text{with probability } 0.1, \end{cases} \tag{3.11}$$

for $i = 1, 2, \ldots, N = 1500$. The table shows standard deviations for $\log(\widehat{\text{fdr}}(z))$, (3.7), from 250 simulations of (3.11), and also using a delta-method formula derived in Section 5 of Efron (2006), incorporated in the *locfdr* algorithm. Rather than (3.6), $f(z)$ was modeled by a seven-parameter natural spline basis, *locfdr*'s default, though this gave nearly the same results as (3.6). Also shown are standard deviations for the corresponding tail-area quantity $\log(\widehat{\text{Fdr}}(z))$ obtained by substituting $\widehat{F}(z) = \int_{-\infty}^{z} \widehat{f}(z')\,dz'$ in (2.3). [This is a little less variable than using $\bar{F}(z)$, (2.4).]

The "Theoretical Null" side of the table shows that $\widehat{\text{fdr}}(z)$ is more variable than $\widehat{\text{Fdr}}(z)$, but both are more than accurate enough for practical use. At $z = 3$, for example, $\widehat{\text{fdr}}(z)$ only errs by about 8%, yielding $\widehat{\text{fdr}}(z) \doteq 0.12 \pm 0.01$. Standard errors are roughly proportional to $N^{-1/2}$, so even reducing N to 250 gives $\widehat{\text{fdr}}(3) \doteq 0.12 \pm .025$, and similarly for other values of z, accurate enough to make pictures like Figure 3 believable.

Empirical Bayes is a bipolar methodology, with alternating episodes of frequentist and Bayesian activity. Frequentists may prefer $\widehat{\text{Fdr}}$ [or $\widehat{Fdr}$, (2.5)] to $\widehat{\text{fdr}}$ because of connections with classical tail-area hypothesis testing, or because cdf's are more straightforward to estimate than densities, while Bayesians prefer $\widehat{\text{fdr}}$ for its more apt a posteriori interpretation. Both, though, combine the Bayesian two-groups model with frequentist estimation methods, and deliver the same basic information.

A variety of local fdr estimation methods have been suggested, using parametric, semiparametric, nonparametric and Bayes methods: Pan et al. (2003), Pounds and Morris (2003), Allison et al. (2002), Heller and Qing (2003), Broberg (2005), Aubert et al. (2004), Liao et al. (2004) and Do et al. (2005), all performing reasonably well. The Poisson GLM methodology of *locfdr* has the advantage of easy implementation with familiar software, and a closed-form error analysis.

Estimation efficiency becomes a more serious problem on the "Empirical Null" side of Table 1, where we can no longer trust the theoretical null $f_0(z) \sim N(0, 1)$. This is the subject of Section 4.

4. THE EMPIRICAL NULL DISTRIBUTION

We have been assuming that $f_0(z)$, the null density in (2.1), is known on theoretical grounds, as in (3.1). This leads to false discovery estimates such as $\widehat{\text{fdr}}(z) = p_0 f_0(z)/\widehat{f}(z)$ and $\widehat{\text{Fdr}}(z) = p_0 F_0(z)/\widehat{F}(z)$, where only denominators need be estimated. Most applications of Benjamini and Hochberg's control algorithm (2.5) make the same assumption (sometimes augmented with permutation calculations, which usually produce only minor corrections to the theoretical null, as discussed in Section 5). Use of the theoretical null is mandatory in classic one-at-a-time testing, where theory provides the only information available for null behavior. But things change in large-scale simultaneous testing situations: serious defects in the theoretical null may become obvious, while empirical Bayes methods can provide more realistic null distributions.

Figure 4 shows z-value histograms for two additional microarray studies, described more fully in Efron (2006). These are of the same form as the prostate data: n subjects in two disease categories provide expression levels for N genes; two-sample t-statistics t_i comparing the categories are computed for each gene, and then transformed to z-values $z_i = \Phi^{-1}(F_{n-2}(t_i))$, as in (3.2). Unlike panel A of Figure 1, however, neither histogram obeys the theoretical $N(0, 1)$ null near $z = 0$. The BRCA data has a much wider central peak, while the HIV peak is too narrow. The lighter curves in Figure 4 are *empirical null* estimates (Efron, 2004), normal curves fit to the central peak of the z-value histograms. The idea here is simple enough: we make the "zero assumption,"

ZERO ASSUMPTION.

(4.1) Most of the z-values near 0 come from null genes,

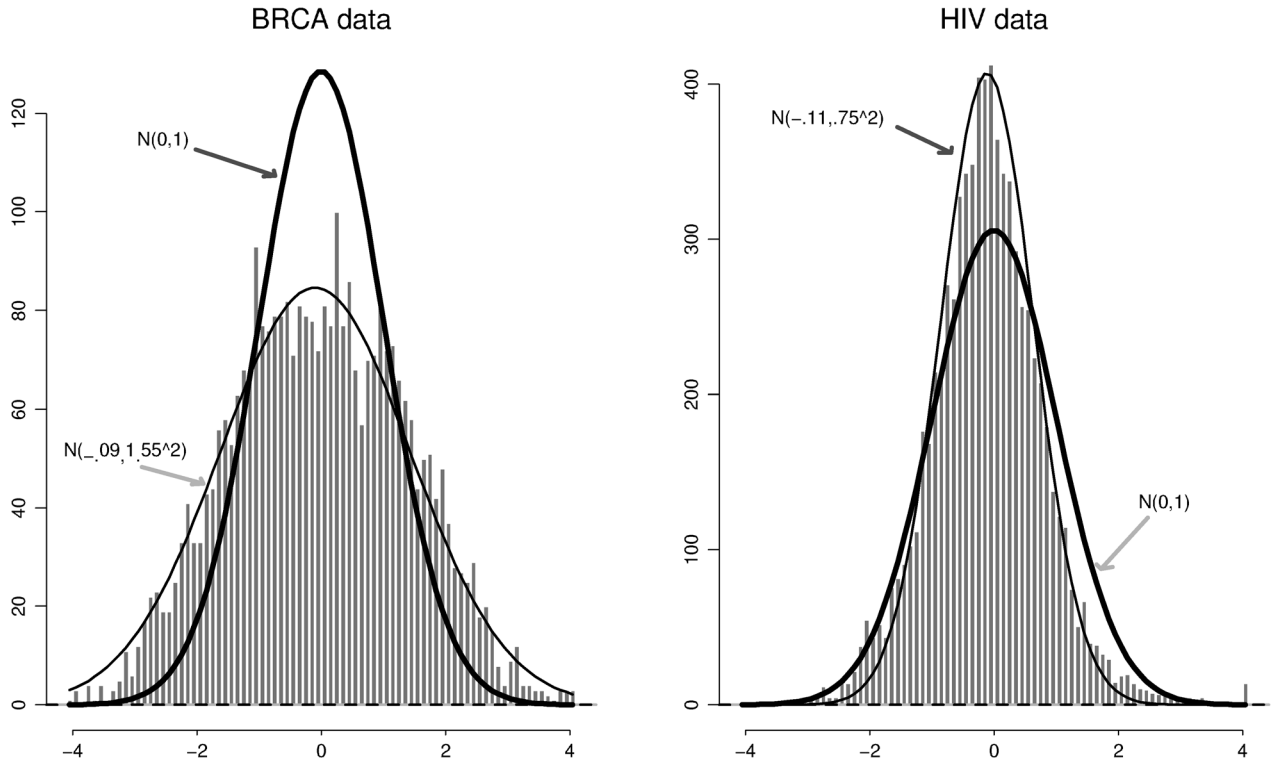

FIG. 4. *z-values from two microarray studies. BRCA data (Hedenfalk et al., 2001), comparing seven breast cancer patients having BRCA1 mutation to eight with BRCA2 mutation $N = 3226$ genes. HIV data (van't Wout et al., 2003) comparing four HIV+ males with four HIV− males, $N = 7680$ genes. Theoretical $N(0, 1)$ null, heavy curve is too narrow for BRCA data, too wide for HIV data. Light curves are empirical nulls: normal densities fit to the central histogram counts.*

(discussed further below), generalize the $N(0, 1)$ theoretical null to $N(\delta_0, \sigma_0^2)$, and estimate (δ_0, σ_0^2) from the histogram counts near $z = 0$. *Locfdr* uses two different estimation methods, analytical and geometric, described next.

Figure 5 shows the geometric method in action on the HIV data. The heavy solid curve is $\log \widehat{f}(z)$, fit from (3.6) using Lindsey's method, as described in Efron and Tibshirani (1996). The two-groups model and the zero assumption suggest that if f_0 is normal, $f(z)$ should be well-approximated near $z = 0$ by $p_0\varphi_{\delta_0,\sigma_0}(z)$, with

$$\varphi_{\delta_0,\sigma_0}(z) \equiv (2\pi\sigma_0^2)^{-1/2}\exp\left\{-\frac{1}{2}\left(\frac{z-\delta_0}{\sigma_0}\right)^2\right\}, \tag{4.2}$$

making $\log f(z)$ approximately quadratic,

$$\begin{aligned}\log f(z) \doteq \log p_0 - \frac{1}{2}\left\{\frac{\delta_0^2}{\sigma_0^2} + \log(2\pi\sigma_0^2)\right\} \\ + \frac{\delta_0}{\sigma_0^2}z - \frac{1}{2\sigma_0^2}z^2.\end{aligned} \tag{4.3}$$

The beaded curve shows the best quadratic approximation to $\log \widehat{f}(z)$ near 0. Matching its coefficients $(\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\beta}_2)$ to (4.3) yields estimates $(\widehat{\delta}_0, \widehat{\sigma}_0, \widehat{p}_0)$, for instance, $\widehat{\sigma}_0 = (2\widehat{\beta}_2)^{-1/2}$,

$$\begin{aligned}\widehat{\delta}_0 &= -0.107,\\ \widehat{\sigma}_0 &= 0.753,\\ \widehat{p}_0 &= 0.931,\end{aligned} \tag{4.4}$$

for the HIV data. Trying the same method with the theoretical null, that is, taking $(\delta_0, \sigma_0) = (0, 1)$ in (4.3), gives a very poor fit, and $\widehat{p}_0$ equals the impossible value 1.20.

The analytic method makes more explicit use of the zero assumption, stipulating that the nonnull density $f_1(z)$ in the two-groups model (2.1) is supported outside some given interval $[a, b]$ containing zero (actually chosen by preliminary calculations). Let N_0 be the number of z_i in $[a, b]$, and define

$$\begin{aligned}P_0(\delta_0, \sigma_0) = \Phi\left(\frac{b-\delta_0}{\sigma_0}\right) - \Phi\left(\frac{a-\delta_0}{\sigma_0}\right) \quad \text{and} \\ \theta = p_0 P_0.\end{aligned} \tag{4.5}$$

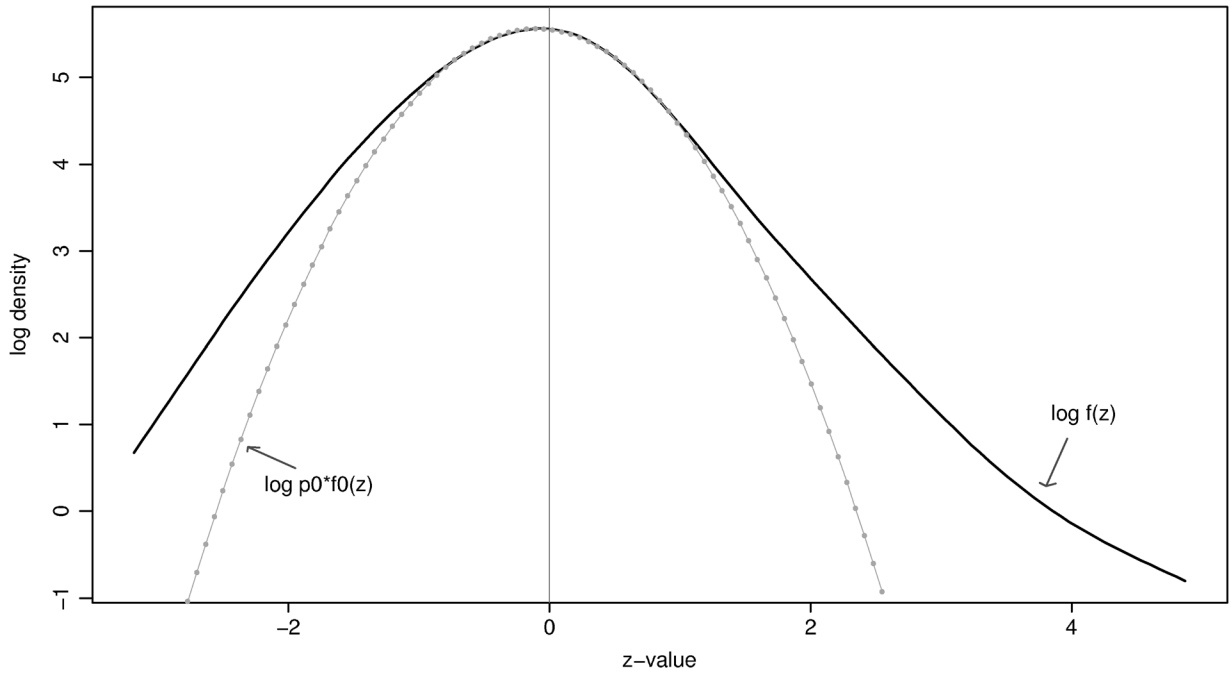

FIG. 5. *Geometric estimate of null proportion* p_0 *and empirical null mean and standard deviation* (δ_0, σ_0) *for the HIV data. Heavy curve is* $\log \widehat{f}(z)$, *estimated as in* (3.3)–(3.6); *beaded curve is best quadratic approximation to* $\log \widehat{f}(z)$ *near* $z = 0$.

Then the likelihood function for $\mathbf{z}_0$, the vector of N_0 z-values in $[a, b]$, is

$$f_{\delta_0,\sigma_0,p_0}(\mathbf{z}_0) = [\theta^{N_0}(1-\theta)^{N-N_0}] \cdot \left[\prod_{z_i \in \mathbf{z}_0} \frac{\varphi_{\delta_0,\sigma_0}(z_i)}{P_0(\delta_0,\sigma_0)}\right]. \tag{4.6}$$

This is the product of two exponential family likelihoods, which is numerically easy to solve for the maximum likelihood estimates $(\widehat{\delta}_0, \widehat{\sigma}_0, \widehat{p}_0)$, equaling $(-0.120, 0.787, 0.956)$ for the HIV data.

Both methods are implemented in *locfdr*. The analytic method is somewhat more stable but can be more biased than geometric fitting. Efron (2004) shows that geometric fitting gives nearly unbiased estimates of δ_0 and σ_0 for $p_0 \geq 0.90$. Table 2 shows how the two methods fared in the simulation study of Table 1.

A healthy literature has sprung up on the estimation of p_0, as in Pawitan et al. (2005) and Langlass et al. (2005), all of which assumes the validity of the theoretical null. The zero assumption plays a central role in this literature [which mostly works with two-sided p-values rather than z-values, e.g., $p_i = 2(1 - F_{100}(|t_i|))$ in (3.2), making the "zero region" occur near $p = 1$]. *The two-groups model is unidentifiable if* f_0 *is unspecified in* (2.1), since we can redefine f_0 as $f_0 + cf_1$, and p_1 as $p_1 - cp_0$ for any $c \leq p_1/p_0$. With p_1 small, (2.2), and f_1 supposed to yield z_i's far from 0 for the most part, the zero assumption is a reasonable way to impose identifiability on the two-groups model. Section 6 considers the meaning of the null density more carefully, among other things explaining the upward bias of $\widehat{p}_0$ seen in Table 2.

The empirical null is an expensive luxury from the point of view of estimation efficiency. Comparing the right-hand side of Table 1 with the left reveals factors of 2 or 3 increase in standard error relative to the theoretical null, near the crucial point where $\text{fdr}(z) = 0.2$. Section 4 of Efron (2005) pins the increased variability entirely on the estimation of (δ_0, σ_0); even knowing the true values of p_0 and $f(z)$ would reduce the standard error of $\log \widehat{\text{fdr}}(z)$ by less than 1%. (Using tail-

TABLE 2

Comparison of estimates $(\widehat{\delta}_0, \widehat{\sigma}_0, \widehat{p}_0)$, *simulation study of Table 1. "Formula" is average from delta-method standard deviation formulas, Section 5 in Efron* (2006), *as implemented in locfdr*

	Geometric			Analytic		
	mean	**stdev**	**(formula)**	**mean**	**stdev**	**(formula)**
$\widehat{\delta}_0$:	0.02	**0.056**	(0.062)	0.04	**0.031**	(0.032)
$\widehat{\sigma}_0$:	1.02	**0.029**	(0.033)	1.04	**0.031**	(0.031)
$\widehat{p}_0$:	0.92	**0.013**	(0.015)	0.93	**0.009**	(0.011)

TABLE 3
Number of genes identified as true discoveries by two-sided Benjamini–Hochberg procedure, 0.10 *control level*

	Theoretical null	Empirical null
BRCA data:	107	0
HIV data:	22	180

Empirical null densities as in Figure 4.

area Fdr's rather than local fdr's does not help—here the local version is less variable.)

The reason for considering empirical nulls is that the theoretical $N(0, 1)$ null does not seem to fit the data in situations like Figure 4. For the BRCA data we can see that the histogram is overdispersed compared to $N(0, 1)$ around $z = 0$; the implication is that there will be more null counts far from zero than the theoretical null predicts, making $N(0, 1)$ false discovery rate calculations like (3.10) too optimistic. The opposite happens with the HIV data.

There is a lot at stake here for both Bayesians and frequentists. Table 3 shows the number of gene discoveries identified by the standard Benjamini–Hochberg two-sided Fdr procedure, $q = 0.10$ in (2.5). The HIV results are much more dramatic using the empirical null $f_0(z) \sim N(-0.11, 0.75^2)$ and in fact we will see in the next section that $\sigma_0 = 0.75$ is quite believable in this case. The BRCA data has been used in the microarray literature to compare analysis techniques, under the presumption that better techniques will produce more discoveries; recently, for instance, in Storey et al. (2005) and Pawitan et al. (2005). Table 3 suggests caution in this interpretation, where using the empirical null negates any discoveries at all.

The z-values in panel C of Figure 1, proteomics data, were calculated from standard Cox likelihood tests that should yield $N(0, 1)$ null results asymptotically. A $N(-0.02, 1.29^2)$ empirical null was obtained from the analytic method, resulting in only one peak with $\widehat{\text{fdr}} < 0.2$; using the theoretical null gave six such peaks.

In panel B of Figure 1, the z-values were obtained from familiar binomial calculations, each z_i being calculated as

$$z = (\widehat{p}_{\text{ad}} - \widehat{p}_{\text{dis}} - \Delta) \cdot \left(\frac{\widehat{p}_{\text{ad}}(1 - \widehat{p}_{\text{ad}})}{n_{\text{ad}}} + \frac{\widehat{p}_{\text{dis}}(1 - \widehat{p}_{\text{dis}})}{n_{\text{dis}}} \right)^{-1/2}, \tag{4.7}$$

where n_{ad} was the number of advantaged students in the high school, $\widehat{p}_{\text{ad}}$ the proportion passing the test, and likewise n_{dis} and $\widehat{p}_{\text{dis}}$ for the disadvantaged students; $\Delta = 0.192$ was the overall difference, median $(\widehat{p}_{\text{ad}})$ − median $(\widehat{p}_{\text{dis}})$. Here the empirical null standard deviation $\widehat{\sigma}_0$ equals 1.52, half again bigger than the theoretical standard deviation we would use if we had only one school's data. An empirical null fdr analysis yielded 75 schools with $\widehat{\text{fdr}} < 0.20$, 30 on the left and 45 on the right. Example B is discussed a bit further in the next two sections, where its use in the two-groups model is questioned.

My point here is not that the empirical null is always the correct choice. The opposite advice, always use the theoretical null, has been inculcated by a century of classic one-case-at-a-time testing to the point where it is almost subliminal, but it exposes the statistician to obvious criticism in situations like the BRCA and HIV data. Large-scale simultaneous testing produces mass information of a Bayesian nature that impinges on individual decisions. The two-groups model helps bring this information to bear, after one decides on the proper choice of f_0 in (2.1). Section 5 discusses this choice, in the form of a list of reasons why the theoretical null, and its close friend the permutation null, might go astray.

5. THEORETICAL, PERMUTATION AND EMPIRICAL NULL DISTRIBUTIONS

Like most statisticians, I have spent my professional life happily testing hypotheses against theoretical null distributions. It came as somewhat of a shock then, when pictures like Figure 4 suggested that the theoretical null might be more theoretical than I had supposed. Once suspicious, it becomes easy to think of reasons why $f_0(z)$, the crucial element in the two-groups model (2.1), might not obey classical guidelines. This section presents four reasons why the theoretical null might fail, and also gives me a chance to say something about the strengths and weaknesses of permutation null distributions.

REASON 1 (*Failed mathematical assumptions*). The usual derivation of the null hypothesis distribution for a two-sample t-statistic assumes independent and identically distributed (i.i.d.) normal components. For the BRCA data of Figure 4, direct inspection of the 3226 by 15 matrix "X" of expression values reveals markedly nonnormal components, skewed to the right (even after the columns of X have been standardized to mean 0 and standard deviation 1, as in all my examples here). Is this causing the failure of the $N(0, 1)$ theoretical null?

Permutation techniques offer quick relief from such concerns. The columns of X are randomly permuted, giving a matrix X^* with corresponding t-values t_i^* and z-values $z_i^* = \Phi^{-1}(F_{n-2}(t_i^*))$. This is done some large number of times, perhaps 100, and the empirical distribution of the $100 \cdot N$ z_i^*'s used as a *permutation null*. The well-known SAM algorithm (Tusher, Tibshirani and Chu, 2001) effectively employs the permutation null cdf in the numerator of the Fdr formula (2.3).

Applied to the BRCA matrix, the permutation null came out nearly $N(0, 1)$ (as did simply simulating the entries of X^* by independent draws from all $3226 \cdot 15$ entries of X), so nonnormal distributions were not the cause of BRCA's overwide histogram. In practice the permutation null usually approximates the theoretical null closely, as a long history of research on the permutation t-test demonstrated; see Section 5.9 of Lehmann and Romano (2005).

REASON 2 (*Unobserved covariates*). The BRCA study is observational rather than experimental—the 15 women were *observed* to be BRCA1 or BRCA2, not *assigned*, and likewise with the HIV and prostate studies. There are likely to be covariates—age, race, general health—that affect the microarray expression levels differently for different genes. If these were known to us, they could be factored out using a separate linear model on each gene's data, providing a new and improved z_i obtained from the "Treatment" coefficient in the model. This would reduce the spread of the z-value histogram, perhaps even restoring the $N(0, 1)$ theoretical null for the BRCA data.

Unobserved covariates act to broaden the null distribution $f_0(z)$. They also broaden the nonnull distribution $f_1(z)$ in (2.1), and the mixture density $f(z)$, but this does not correct fdr estimates like (3.10), where the numerator, which depends entirely on f_0, is the only estimated quantity. Section 4 of Efron (2004) provides an analysis of a simplified model with unobserved covariates. Permutation techniques cannot recognize unobserved covariates, as the model demonstrates.

REASON 3 (*Correlation across arrays*). False discovery rate methodology does not require independence among the test statistics z_i. However, the theoretical null distribution does require independence of the expression values used to calculate each z_i; in terms of the elements x_{ij} of the expression matrix X, for gene i we need independence among $x_{i1}, x_{i2}, \ldots, x_{in}$ in order to validate (1.1).

Experimental difficulties can undercut across-microarray independence, while remaining undetectable in a permutation analysis. This happened in both studies of Figure 4 (Efron, 2004, 2006). The BRCA data showed strong positive correlations among the first four BRCA2 arrays, and also among the last four. This reduces the effective degrees of freedom for each t-statistic below the nominal 13, making t_i and $z_i = \Phi^{-1}(F_{13}(t_i))$ overdispersed.

REASON 4 (*Correlation across genes*). Benjamini and Hochberg's 1995 paper verified Fdr control for rule (2.5) under the assumption of independence among the N z-values (relaxed a little in Benjamini and Yekutieli, 2001). This seems fatal for microarray applications since we expect genes to be correlated in their actions. A great virtue of the empirical Bayes/two-groups approach is that independence is not necessary; with $\widehat{\mathrm{Fdr}}(z) = p_0 F_0(z)/\widehat{F}(z)$, for instance, $\widehat{\mathrm{Fdr}}(z)$ can provide a reasonable estimate of $\Pr\{\text{null}|Z \le z\}$ as long as $\widehat{F}(z)$ is roughly unbiased for $F(z)$—in formal terms requiring consistency but not independence—and likewise for the local version $\widehat{\mathrm{fdr}}(z) = p_0 f_0(z)/\widehat{f}(z)$, (3.7).

There is, however, a black cloud inside the silver lining: the assumption that the null density $f_0(z)$ is known to the statistician. The empirical null estimation methods of Section 4 do not require z-value independence, and so disperse the black cloud, at the expense of increased variability in fdr estimates. Do we really need to use an empirical null? Efron (2007) discusses the following somewhat disconcerting result: even if the theoretical null distribution $z_i \sim N(0, 1)$ holds exactly true for all null genes, Reasons 1–3 above not causing trouble, correlation among the z_i's can make the overall null distribution effectively much wider or much narrower than $N(0, 1)$.

Microarray data sets tend to have substantial z-value correlations. Consider the BRCA data: there are more than five million correlations ρ_{ij} between pairs of gene z-values z_i and z_j; by examining the row-wise correlations in the X matrix we can estimate that the distribution of the ρ_{ij}'s has approximately mean 0 and variance $\alpha^2 = 0.153^2$,

$$\rho \sim (0, \alpha^2). \tag{5.1}$$

(The zero mean is a consequence of standardizing the columns of X.) This is a lot of correlation—as much as if the BRCA genes occurred in 10 independent groups, but with common interclass correlation 0.50 for all genes within a group.

Section 3 of Efron (2006) shows that under assumptions (1.1)–(5.1), the ensemble of null-gene z-values will behave roughly as

$$z_i \dot{\sim} N(0, \sigma_0^2) \tag{5.2}$$

with

$$\sigma_0^2 = 1 + \sqrt{2}A, \quad A \sim (0, \alpha^2). \tag{5.3}$$

If the variable A equaled $\alpha = 0.153$, for instance, giving $\sigma_0 = 1.10$, then the expected number of null counts below $z = -3$ would be about $p_0 N \Phi(-3/1.10)$ rather than $p_0 N \Phi(-3)$, more than twice as many. There is even more correlation in the HIV data, $\alpha \doteq 0.42$, enough so that a moderately negative value of A could cause $\sigma_0 = 0.75$, as in Figure 4.

The random variable A acts like an observable ancillary in the two-groups situation—observable because we can estimate σ_0 from the central counts of the z-value histogram, as in Section 4; $\widehat{\sigma}_0$ is essentially the half-width of the central peak.

Figure 6 is a cautionary story on the dangers of ignoring $\widehat{\sigma}_0$. A simulation model with

$$\begin{aligned} z_i &\sim N(0,1), \quad i = 1, 2, \ldots, 2700, \quad \text{and} \\ z_i &\sim N(2.5, 1.5), \quad i = 2701, \ldots, 3000, \end{aligned} \tag{5.4}$$

was run, in which the null z_i's, the first 2700, were correlated to the same degree as in the BRCA data, $\alpha = 0.153$. For each of 1000 simulations of (5.4), a standard Benjamini–Hochberg Fdr analysis (2.5) (i.e., using the theoretical null for F_0) was run at control level $q = 0.10$, and used to identify a set of nonnull genes.

Each of the thousand points in Figure 6 is $(\widehat{\sigma}_0, \text{Fdp})$, where $\widehat{\sigma}_0$ is half the distance between the 16th and 86th percentiles of the 3000 z_i's, and Fdp is the "False discovery proportion," the proportion of identified genes that were actually null. Fdp averaged 0.091, close to the target value $q = 0.10$, but with a strong dependence on $\widehat{\sigma}_0$: the lowest 5% of $\widehat{\sigma}_0$'s corresponded to Fdp's averaging only 0.03, while the upper 5% average was 0.29, a factor of 9 difference.

The point here is not that the claimed q-value 0.10 is wrong, but that in any one simulation we may be able to see, from $\widehat{\sigma}_0$, that it is probably misleading. Using the empirical null counteracts this fallacy which, again, is not apparent from the permutation null. (Section 4 of Efron, 2007, discusses more elaborate permutation methods that do bear on Figure 6. See Qui et al., 2005, for a gloomier assessment of correlation effects in microarray analyses.)

What is causing the overdispersion in the Education data of panel B, (4.7)? Correlation across schools, Reason 44, seems ruled out by the nature of the sampling, leaving Reasons 2 and 3 as likely candidates; un-

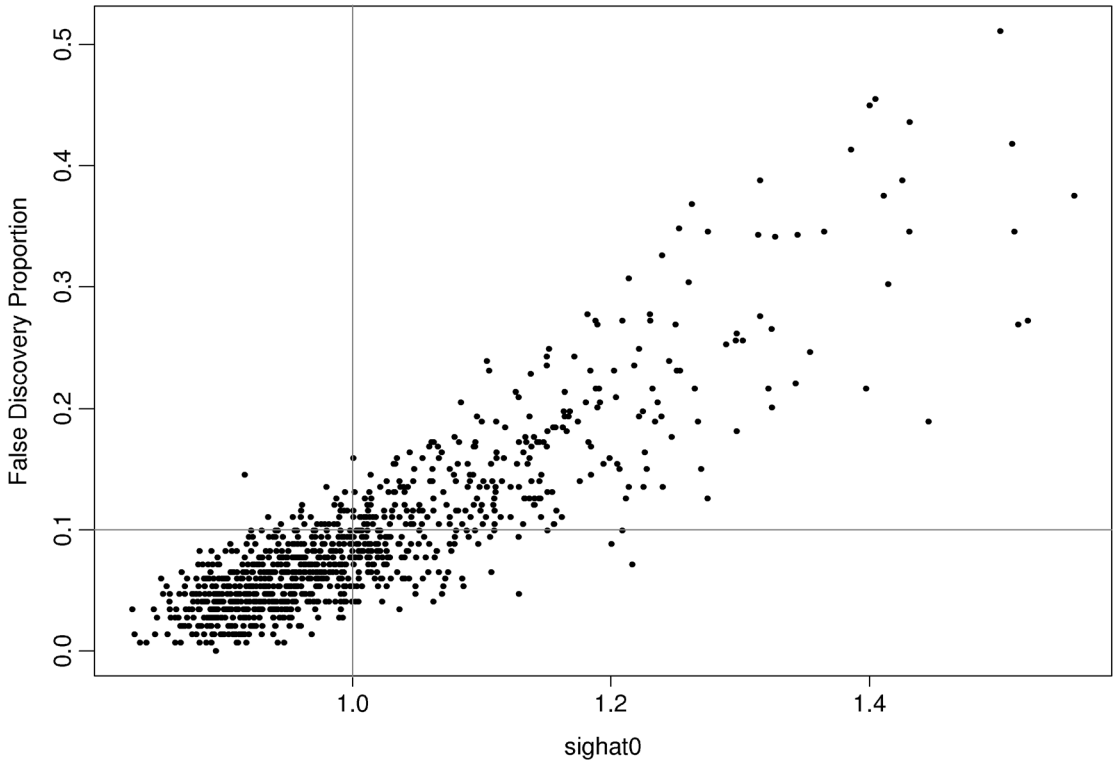

FIG. 6. *Benjamini–Hochberg Fdr control procedure* (2.5), $q = 0.1$, *run for* 1000 *simulations of correlated model* (5.4); *true false discovery proportion Fdp plotted versus half-width estimate* $\widehat{\sigma}_0$. *Overall Fdp averaged* 0.091, *close to* q, *but with a strong dependence on* $\widehat{\sigma}_0$.

observed covariates are an obvious threat here, while within-school sampling dependences (Reason 3) are certainly possible. Fdr analysis yields eight times as many "significant" schools based on the theoretical null rather than $f_0 \sim N(-0.35, 1.51^2)$, but looks completely untrustworthy to me.

Sometimes the theoretical null distribution is fine, of course. The prostate data had $(\widehat{\delta}_0, \widehat{\sigma}_0) = (0.00, 1.06)$ according to the analytic method of (4.6), close enough to $(0, 1)$ to make theoretical null calculations believable. However, there are lots of things that can go wrong with the theoretical null, and lots of data to check it with in large-scale testing situations, making it a matter of due diligence for the statistician to do such checking, even if only by visual inspection of the z-value histogram. All simultaneous testing procedures, not just false discovery rates, go wrong if the null distribution is misrepresented.

6. A ONE-GROUP MODEL

Classical one-at-a-time hypothesis testing depends on having a unique null density $f_0(z)$, such as Student's t distribution for the normal two-sample situation. The assumption of unique f_0 has been carried over into most of the microarray testing literature, including our definition (2.1) of the two-groups model.

Realistic examples of large-scale inference are apt to be less clearcut, with true effect sizes ranging continuously from zero or near zero to very large. Here we consider a "one-group" structural model that allows for a range of effects. We can still usefully apply fdr methods to data from one-group models; doing so helps clarify the choice between theoretical and empirical null hypotheses, and explicates the biases inherent in model (2.1). The discussion in this section, as in Section 2, will be mostly theoretical, involving probability models rather than collections of observed z-values.

Model (2.1) does not require knowing how the z-values were generated, a substantial practical advantage of the two-groups formulation. In contrast, one-group analysis begins with a specific Bayesian structural model. We assume that the ith case has an unobserved *true value* μ_i distributed according to some density $g(\mu)$, and that the observed z_i is normally distributed around μ_i,

$$\mu \sim g(\cdot) \quad \text{and} \quad z|\mu \sim N(\mu, 1). \tag{6.1}$$

The density $g(\mu)$ is allowed to have discrete atoms. It might have an atom at zero but this is not required, and in any case there is no a priori partition of $g(\mu)$ into null and nonnull components.

As an example, suppose $g(\mu)$ is a mixture of 90% $N(0, 0.5^2)$ and 10% $N(2.5, 0.5^2)$,

$$g(\mu) = 0.9 \cdot \varphi_{0,0.5}(\mu) + 0.1 \cdot \varphi_{2.5,0.5}(\mu) \tag{6.2}$$

in notation (4.2). The histogram in Figure 7 shows $N = 3000$ draws of μ_i from (6.2). I am thinking of this as a situation having a large proportion of uninteresting cases centered near, but not exactly at, zero, and a small proportion of interesting cases centered far to the right. We still want to use the observed z_i's from (6.2) to flag cases that are likely to be interesting.

The density of z in model (6.1) is

$$f(z) = \int_{-\infty}^{\infty} \varphi(\mu - z) g(\mu)\, d\mu, \qquad [\varphi(x) = \exp(-x^2/2)/\sqrt{2\pi}], \tag{6.3}$$

shown as the smooth curve in the left-hand panel,

$$f(z) = 0.9 \cdot \varphi_{0,1.12}(z) + 0.1 \cdot \varphi_{2.5,1.12}(z). \tag{6.4}$$

The effect of noise in going from μ_i to $z_i \sim N(\mu_i, 1)$ has blurred the strongly bimodal μ-histogram into a smoothly unimodal $f(z)$.

We can still employ the tactic of Figure 5, fitting a quadratic curve to $\log f(z)$ around $z = 0$ to estimate p_0 and the empirical null density $f_0(z)$. Using the formulas described later in this section gives

$$p_0 = 0.93 \quad \text{and} \quad f_0(z) \sim N(.02, 1.14^2), \tag{6.5}$$

and corresponding fdr curve $p_0 f_0(z)/f(z)$, labeled "Emp null" in the right-hand panel of Figure 7.

Looking at the histogram, it is reasonable to consider "interesting" those cases with $\mu_i \geq 1.5$, and "uninteresting" $\mu_i < 1.5$. The curve labeled "Bayes" in Figure 7 is the posterior probability $\Pr\{\text{uninteresting}|z\}$ based on full knowledge of (6.1), (6.2). The empirical null fdr curve provides an excellent estimate of the full Bayes result, without the prior knowledge. [An fdr based on the theoretical $N(0, 1)$ null is seen to be far off.]

Unobserved covariates, Reason 2 in Section 4, can easily produce blurry null hypotheses like that in (6.2). My point here is that the two-group model will handle blurry situations if the null hypothesis is empirically estimated. Or, to put things negatively, theoretical or permutation null methods are prone to error in such situations, no matter what kind of analysis technique is used.

Comparing (6.5) with (6.4) shows that $f_0(z)$ is just about right, but p_0 is substantially larger than the value 0.90 we might expect. The $\varphi_{2.5,.5}$ component of $g(\mu)$

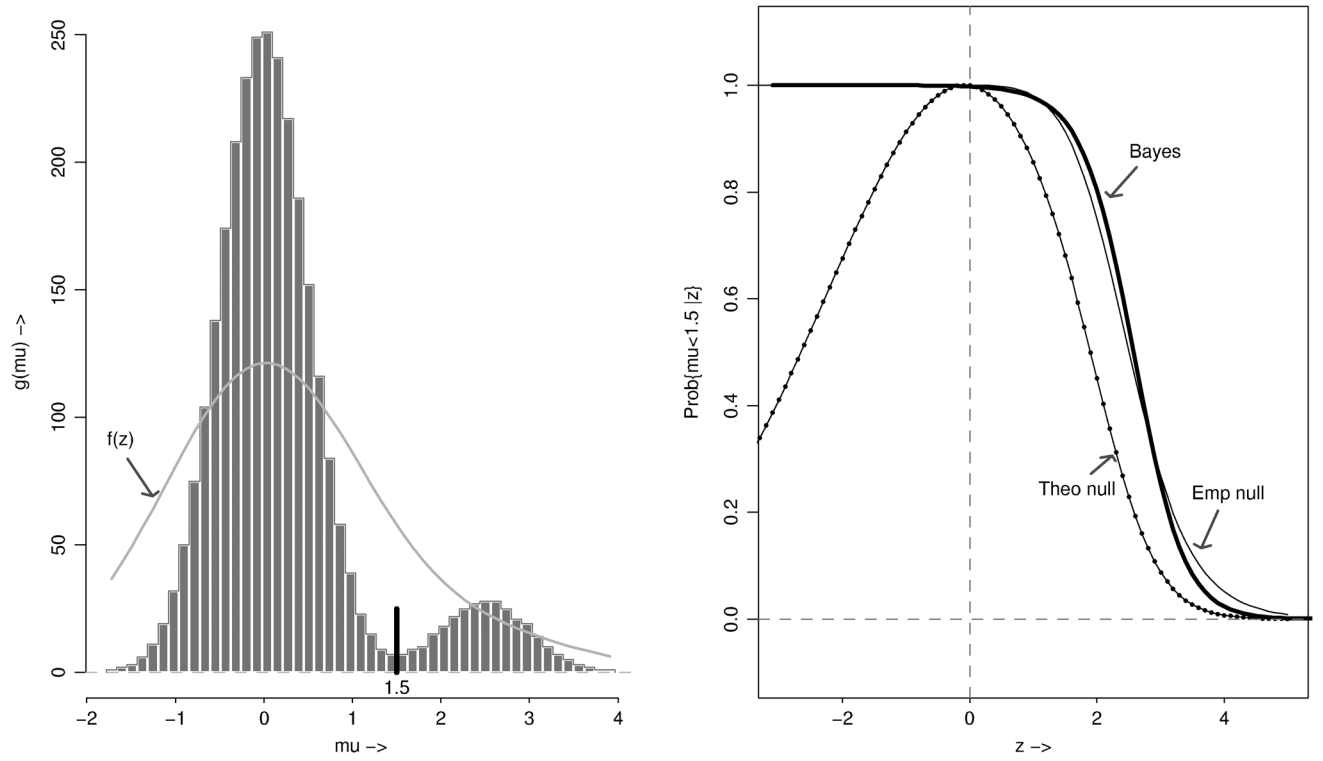

FIG. 7. *Left panel*: *Histogram shows* $N = 3000$ *draws of* μ_i *from model* (6.2); *smooth curve is corresponding density* $f(z)$, (6.3). *Right panel*: "*Emp null*" *is fdr*(z) *based on empirical null*; *it closely matches full Bayes posterior probability* "*Bayes*" $= \Pr\{\mu_k < 1.5|z\}$ *from* (6.1)–(6.2); "*Theo null*" *is fdr*(z) *based on theoretical null*, *a poor match to Bayes*.

puts some of its z-values near zero, weakening the zero assumption (4.1) and biasing p_0 upward. The same thing happened in Table 2 even though model (3.11) is "unblurred," $g(\mu)$ having a point mass at $\mu = 0$. Fortunately, p_0 is the least important part of the two-groups model for estimating fdr(z), under assumption (2.2). "Bias" can be a misleading term in model (6.1) since it presupposes that each μ_i is clearly defined as null or nonnull. This seems clear enough in (3.11). The null/nonnull distinction is less clear in (6.2), though it still makes sense to search for cases that have μ_i unusually far from 0.

The results in (6.5) come from a theoretical analysis of model (6.1). The idea in what follows is to generalize the construction in Figure 5 by approximating $\ell(z) = \log f(z)$ with Taylor series other than quadratic.

The Jth Taylor approximation to $\ell(z)$ is

$$\ell_J(z) = \sum_{j=0}^{J} \ell^{(j)}(0) z^j / j!, \tag{6.6}$$

where $\ell^{(0)}(0) = \log f(0)$ and for $j \geq 1$

$$\ell^{(j)}(0) = \frac{d^j \log f(z)}{dz^j}\bigg|_{z=0}. \tag{6.7}$$

Let $\widetilde{f}_0(z)$ indicate the subdensity $p_0 f_0(z)$, the numerator of fdr(z) in (2.7). The choice

$$\widetilde{f}_0(z) = e^{\ell_J(z)} \tag{6.8}$$

matches $f(z)$ at $z = 0$ (a convenient form of the zero assumption) and leads to an fdr expression

$$\text{fdr}(z) = e^{\ell_J(z)} / f(z). \tag{6.9}$$

Larger choices of J match $\widetilde{f}_0(z)$ more accurately to $f(z)$, increasing ratio (6.9); the interesting z-values, those with small fdr's, are pushed farther away from zero as we allow more of the data structure to be explained by the null density.

Bayesian model (6.1) provides a helpful interpretation of the derivatives $\ell^{(j)}(0)$:

LEMMA. *The derivative* $\ell^{(j)}(0)$, (6.7), *is the* j*th cumulant of the posterior distribution of* μ *given* $z = 0$, *except that* $\ell^{(2)}(0)$ *is the second cumulant minus* 1. *Thus*

$$\begin{aligned} \ell^{(1)}(0) &= E_0 \quad \textit{and} \\ -\ell^{(2)}(0) &= 1 - V_0 \equiv \bar{V}_0, \end{aligned} \tag{6.10}$$

where E_0 *and* V_0 *are the posterior mean and variance of* μ *given* $z = 0$.

TABLE 4
Expressions for p_0, f_0 and fdr, first three choices of J in (6.8), (6.9); $\bar{V}_0 = 1 - V_0$; $J = 0$ gives theoretical null, $J = 2$ empirical null; $f(z)$ from (6.3)

J	0	1	2
p_0	$f(0)\sqrt{2\pi}$	$f(0)\sqrt{2\pi}e^{E_0^2/2}$	$f(0)\sqrt{\frac{2\pi}{\bar{V}_0}}e^{E_0^2/2\bar{V}_0}$
$f_0(z)$	$N(0,1)$	$N(E_0,1)$	$N(E_0/\bar{V}_0, 1/\bar{V}_0)$
fdr(z)	$\frac{f(0)e^{-z^2/2}}{f(z)}$	$\frac{f(0)e^{E_0 z - z^2/2}}{f(z)}$	$\frac{f(0)e^{E_0 z - \bar{V}_0 z^2/2}}{f(z)}$

Proof of the lemma appears in Section 7 of Efron (2005).

For $J = 0, 1, 2$, formulas (6.8), (6.9) yield simple expressions for p_0 and $f_0(z)$ in terms of $f(0)$, E_0 and $\bar{V}_0$. These are summarized in Table 4, with p_0 obtained from

$$p_0 = \int_{-\infty}^{\infty} \tilde{f}_0(z)\,dz. \tag{6.11}$$

Formulas are also available for Fdr(z), (2.8).

The choices $J = 0, 1, 2$ in Table 4 result in a normal null density $f_0(z)$, the only difference being the means and variances. Going to $J = 3$ allows for an asymmetric choice of $f_0(z)$,

$$\text{fdr}(z) = \frac{f(0)}{f(z)} e^{E_0 z - \bar{V}_0 z^2/2 + S_0 z^3/6}, \tag{6.12}$$

where S_0 is the posterior third central moment of μ given $z = 0$ in model (6.1). The program *locfdr* uses a variant, the "split normal," to model asymmetric null densities, with the exponent of (6.12) replaced by a quadratic spline in z.

The lemma bears on the difference between empirical and theoretical nulls. Suppose that the probability mass of $g(\mu)$ occurring within a few units of the origin is concentrated in an atom at $\mu = 0$. Then the posterior mean and variance (E_0, V_0) of μ given $z = 0$ will be near 0, making $(E_0, \bar{V}_0) \doteq (0, 1)$. In this case the empirical null ($J = 2$) will approximate the theoretical null ($J = 0$). Otherwise the two nulls differ; in particular, any mass of $g(\mu)$ near zero increases V_0, swelling the standard deviation $(1 - V_0)^{-1/2}$ of the empirical null.

The two-groups model (2.1), (2.2) puts one in a hypothesis-testing frame of mind: a large group of uninteresting cases is to be statistically separated from a small interesting group. Even blurry situations like (6.2) exhibit a clear grouping, as in Figure 7. None of this is necessary for the one-group model (6.1). We might, for example, suppose that $g(\mu)$ is normal,

$$\mu \sim N(A, B^2), \tag{6.13}$$

and proceed in an empirical Bayes way to estimate A and B and then apply Bayes estimation to the individual cases.

This line of thought leads directly to James–Stein estimation (Efron and Morris, 1975). Estimation, as opposed to testing, is the key word here—with possible effect sizes μ_i varying continuously rather than having a large clump of values near zero. The Education data of panel B, Figure 1, could reasonably be analyzed this way, instead of through simultaneous testing. Scientific context, which says that there is likely to be a large group of (nearly) unaffected genes, as in (2.2), is what makes the two-groups model a reasonable Bayes prior for microarray studies.

7. BAYESIAN AND FREQUENTIST CONFIDENCE STATEMENTS

False discovery rate methods provide a happy marriage between Bayesian and frequentist approaches to multiple testing, as shown in Section 2. Empirical Bayes techniques based on the two-groups model seem to give us the best of both statistical philosophies. Things do not always work out so peaceably; in these next two sections I want to discuss contentious situations where the divorce court looms as a possibility.

An insightful and ingenious paper by Benjamini and Yekutieli (2005) discusses the following problem in simultaneous significance testing: having applied false discovery rate methods to select a set of nonnull cases, how can confidence intervals be assigned to the true effect size for each selected case? (The paper and the ensuing discussion are much more general, but this is all I need for the illustration here.)

Figure 8 concerns Benjamini and Yekutieli's solution applied to the following simulated data set: $N = 10{,}000$ (μ_i, z_i) pairs were generated as in (6.1), with 90% of the μ_i zero, the null cases, and 10% distributed $N(-3, 1)$,

$$g(\mu) = 0.90 \cdot \delta_0(\mu) + 0.10 \cdot \varphi_{-3,1}(\mu), \tag{7.1}$$

$\delta_0(\mu)$ a delta function at $\mu = 0$. The Fdr procedure (2.5) was applied with $q_0 = 0.05$, yielding 566 nonnull "discoveries," those having $z_i \le -2.77$.

The Benjamini–Yekutieli "false coverage rate" (FCR) control procedure provides upper and lower bounds for the true effect size μ_i corresponding to each z_i less than -2.77; these are indicated by heavy diagonal lines in Figure 8, constructed as described in BY's Definition 1. This construction guarantees that

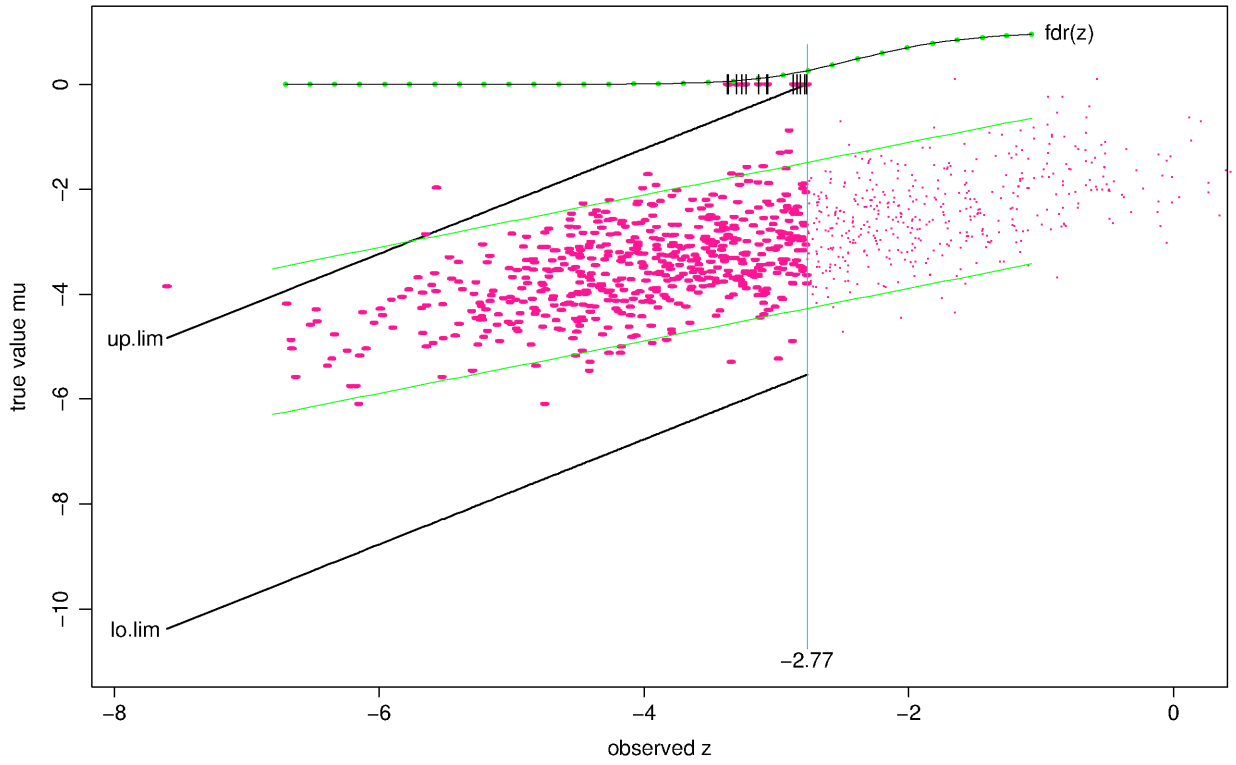

FIG. 8. *Benjamini–Yekutieli FCR controlling intervals applied to simulated sample of* 10,000 *cases from* (6.1), (7.1). 566 *cases have* $z_i \le z_0 = -2.77$, *the Fdr* (0.05) *threshold. Plotted points are* (z_i, μ_i) *for the* 1000 *nonnull cases*; 14 *null cases with* $z_i \le z_0$ *indicated by* "+." *Heavy diagonal lines indicate FCR* 95% *interval limits*; *light lines are Bayes* 95% *posterior intervals given* $\mu_i \ne 0$. *Beaded curve at top is fdr*(z_i), *posterior probability* $\mu_i = 0$.

the expected proportion of the 566 intervals *not* containing the true μ_i, the false coverage rate, is bounded by $q = 0.05$.

In a real application only the z_i's and their BY confidence intervals could be seen, but in a simulation we can plot the actual (z_i, μ_i) pairs, and compare them to the intervals. Figure 8 plots (z_i, μ_i) for the 1000 nonnull cases, those from $\mu_i \sim N(-3, 1)$ in (7.1). Of these, 55,2 plotted as heavy points, lie to the left of $z_0 = -2.77$, the Fdr threshold, with the other 448 plotted as light points; 14 null cases, $\mu_i = 0$, plotted as "+," also had $z_i < z_0$.

The first thing to notice is that the FCR property is satisfied: only 17 of the 566 intervals have failed to contain μ_i (14 of these the +'s), giving 3% noncoverage. The second thing, though, is that the intervals are frighteningly wide—$z_i \pm 2.77$, about $\sqrt{2}$ longer than the usual individual 95% intervals $z_i \pm 1.96$—and poorly centered, particularly at the left where all the μ_i's fall in their intervals' upper halves.

An interesting comparison is with Bayes' rule applied to (6.1), (7.1), which yields

$$\Pr\{\mu = 0 | z_i\} = \text{fdr}(z_i), \tag{7.2}$$

where

$$\text{fdr}(z) = 0.9 \cdot \varphi_{0,1}(z) \cdot [0.9 \cdot \varphi_{0,1}(z) + 0.1 \cdot \varphi_{-3,\sqrt{2}}(z)]^{-1} \tag{7.3}$$

as in (2.7), and

$$g(\mu_i | \mu_i \ne 0, z_i) \sim N\left(\frac{z_i - 3}{2}, \frac{1}{2}\right). \tag{7.4}$$

That is, μ_i is null with probability $\text{fdr}(z_i)$, and $N((z_i - 3)/2,\ 1/2)$ with probability $1 - \text{fdr}(z_i)$. The dashed lines indicate the posterior 95% intervals given that μ_i is nonnull, $(z_i - 3)/2 \pm 1.96/\sqrt{2}$, now $\sqrt{2}$ *shorter* than the usual individual intervals; at the top of Figure 9 the beaded curve shows $\text{fdr}(z_i)$.

The frequentist FCR intervals and the Bayes intervals are pursuing the same goal, to include the nonnull scores μ_i with 95% probability. At $z_i = -2.77$ the FCR assessment is $\Pr\{\mu \in [-5.54, 0]\} = 0.95$; Bayes' rule states that $\mu_i = 0$ with probability $\text{fdr}(-2.77) = 0.25$, and if $\mu_i \ne 0$, then $\mu_i \in [-4.27, -1.49]$ with probability 0.95. This kind of disconnected description is natural to the two-groups model. A principal cause of FCR's oversized intervals (the paper shows

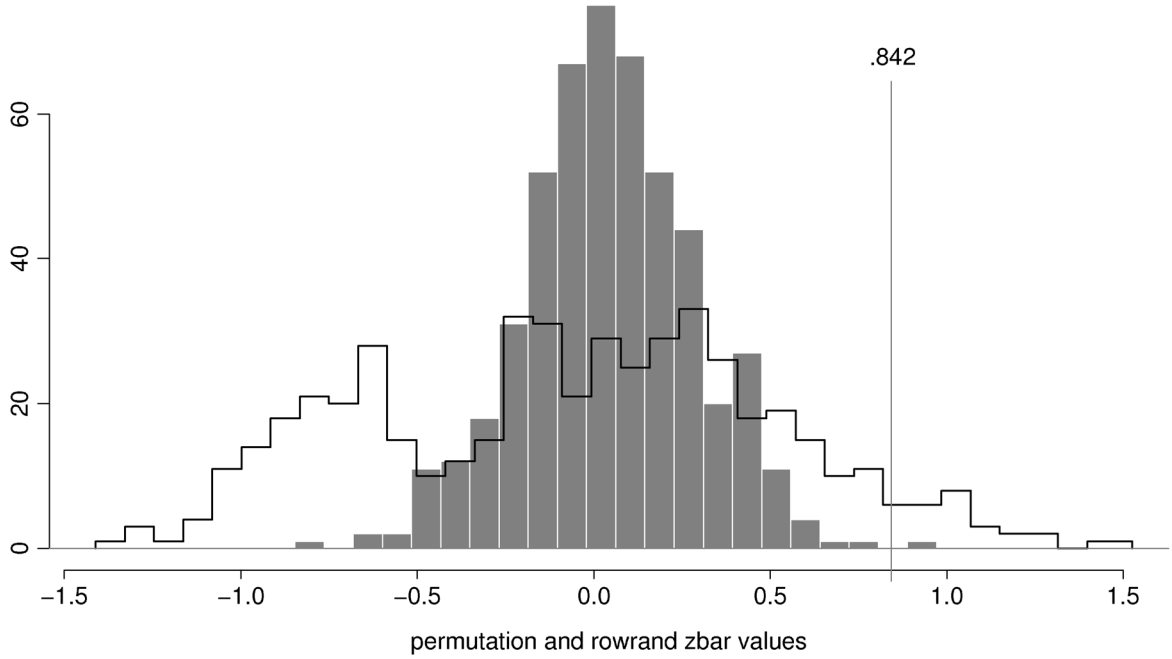

FIG. 9. *Computing a p-value for* $\bar{z}_{\mathcal{S}} = 0.842$, *average of* 15 z*-values in CTL pathway, p*53 *data Solid histogram* 500 *row randomizations give p-value* 0.002. *Line histogram* 500 *column permutations give p-value* 0.048.

that no FCR-controlling intervals can be much narrower) comes from using a single connected set to describe a disconnected situation.

Of course Bayes' rule will not be easily available to us in most practical problems. Is there an empirical Bayes solution? Part of the solution certainly is there: estimating fdr(z) as in Section 3. Estimating $g(\mu_i|\mu_i \neq 0, z_i)$, (7.4), is more challenging. A straightforward approach uses the nonnull counts (3.8) to estimate the nonnull density $f_1(z)$ in (2.1), deconvolutes $\widehat{f}_1(z)$ to estimate the nonnull component "$g_1(\mu)$" in (7.1), and applies Bayes' rule directly to $\widehat{g}_1$. This works reasonably well in Figure 8's example, but deconvolution calculations are notoriously tricky and I have not been able to produce a stable general algorithm.

Good frequentist methods like the FCR procedure enjoy the considerable charm of an exact error bound, without requiring a priori specifications, and of course there is no law that they have to agree with any particular Bayesian analysis. In large-scale situations, however, empirical Bayes information can overwhelm both frequentist and Bayesian predilections, hopefully leading to a more satisfactory compromise between the two sets of intervals appearing in Figure 8.

8. IS A SET OF GENES ENRICHED?

Microarray experiments, through a combination of insufficient data per gene and massively multiple simultaneous inference, often yield disappointing results. In search of greater detection power, *enrichment analysis* considers the combined outcomes of biologically defined sets of genes, such as pathways. As a hypothetical example, if the 20 z-values in a certain pathway all were positive, we might infer significance to the pathway's effect, whether or not any of the individual z_i's were deemed nonnull.

Our example here will involve the *p53 data*, from Subramanian et al. (2005), $N = 10{,}100$ genes on $n = 50$ microarrays, z_i's as in (3.2), whose z-value histogram looks like a slightly short-tailed normal distribution having mean 0.04 and standard deviation 1.06. Fdr analysis (2.5), $q = 0.1$, yielded just one nonnull gene, while enrichment analysis indicated seven or eight significant gene sets, as discussed at length in Efron and Tibshirani (2006).

Figure 9 concerns the *CTL pathway*, a set of 15 genes relating to the development of so-called killer T cells, #95 in a catalogue of 522 gene-sets provided by Subramanian et al. (2005). For a given gene-set "$\mathcal{S}$" with m members, let $\bar{z}_{\mathcal{S}}$ denote the mean of the m z-values within $\mathcal{S}$; $\bar{z}_{\mathcal{S}}$ is the enrichment statistic suggested in the Bioconductor *R* package *limma* (Smyth, 2004),

$$\bar{z}_{\mathcal{S}} = 0.842 \tag{8.1}$$

for the CTL pathway. How significant is this result? I will consider assigning an individual p-value to (8.1),

not taking into account multiple inference for a catalogue of possible gene-sets (which we could correct for later using Fdr methods, for instance, to combine the individual p-values).

Limma computes p-values by "row randomization," that is, by randomizing the order of rows of the $N \times n$ expression matrix X, and recomputing the statistic of interest. For a simple average like (8.1) this amounts to choosing random subsets of size $m = 15$ from the $N = 10,100$ z_i's and comparing $\bar{z}_{\mathcal{S}}$ to the distribution of the randomized values $\bar{z}^*_{\mathcal{S}}$. Five hundred rowrands produced only one $\bar{z}^*_{\mathcal{S}} > \bar{z}_{\mathcal{S}}$, giving p-value $1/500 = 0.002$.

Subramanian et al. calculate p-values by permuting the *columns* of X rather than the rows. The permutations yield a much wider distribution than the row randomizations in Figure 9, with corresponding p-value 0.048. The reason is simple: the genes in the CTL pathway have highly correlated expression levels that increase the variance of $\bar{z}^*_{\mathcal{S}}$; column-wise permutations of X preserve the correlations across genes, while row randomizations destroy them.

At this point it looks like column permutations should always give the right answer. Wrong! For the BRCA data in Figure 4, the ensemble of z-values has (mean, standard deviation) about $(0, 1.50)$, compared to $(0, 1)$ for z_i^*'s from column permutations. This shrinks the permutation variability of $\bar{z}^*_{\mathcal{S}}$, compared to what one would get from a random selection of genes for $\mathcal{S}$, and can easily reverse the relationship in Figure 9.

The trouble here is that there are two obvious, but different, null hypotheses for testing enrichment:

Randomization null hypothesis $\mathcal{S}$ has been chosen by random selection of m genes from the full set of N genes.

Permutation null hypothesis The order of the n microarrays has been chosen at random with respect to the patient characteristics (e.g., with the patient being in the normal or cancer category in Example A of the Introduction).

Efron and Tibshirani (2006) suggest a compromise method, *restandardization*, that to some degree accommodates both null hypotheses. Instead of permuting $\bar{z}_{\mathcal{S}}$ in (8.1), restandardization permutes $(\bar{z}_{\mathcal{S}} - \mu_z)/\sigma_z$, where (μ_z, σ_z) are the mean and standard deviation of all N z_i's. Subramanian et al. do something similar using a Kolmogorov–Smirnov enrichment statistic.

All of these methods are purely frequentistic. Theoretically we might consider applying the two-groups/empirical Bayes approach to sets of z-values "$\mathbf{z}_{\mathcal{S}}$," just as we did for individual z_i's in Sections 2 and 3. For at least three reasons that turns out to be extremely difficult:

- My technique for estimating the mixture density f, as in (3.6), becomes exponentially more difficult in higher dimensions.
- There is not likely to be satisfactory theoretical null f_0 for the correlated components of $\bar{z}_{\mathcal{S}}$, while estimating an empirical null faces the same "curse of dimensionality" as for f.
- As discussed following (3.10), false discovery rate interpretation depends on exchangeability, essentially an equal a priori interest in all N genes. There may be just one gene-set $\mathcal{S}$ of interest to an investigator, or a catalogue of several hundred $\mathcal{S}$'s as in Subramanian et al., but we certainly are not interested in all possible gene-sets. It would be a daunting exercise in subjective, as opposed to empirical, Bayesianism to assign prior probabilities to any particular gene-set $\mathcal{S}$.

Having said this, it turns out there is one "gene-set" situation where the two-groups/empirical Bayes approach is practical (though it does not involve genes). Looking at panel D of Figure 1, the Imaging data, the obvious spatial correlation among z-values suggests local averaging to reduce the effects of noise.

This has been carried out in Figure 10: at voxel i of the $N = 15,445$ voxels, the average of z-values for those voxels within city-block distance 2 has been computed, say "$\bar{z}_i$." The results for the same horizontal slice as in panel D are shown using a similar symbol code. Now that we have a single number z_i for each voxel, we can compute the empirical null $\widehat{\text{fdr}}$ estimates as in Section 4. The voxels labeled "enriched" in Figure 10 are those having $\widehat{\text{fdr}}(\bar{z}_i) \le 0.2$.

Enrichment analysis looks much more familiar in this example, being no more than local spatial smoothing. The convenient geometry of three-dimensional space has come to our rescue, which it emphatically fails to do in the microarray context.

9. CONCLUSION

Three forces influence the state of statistical science at any one time: mathematics, computation and applications, by which I mean the type of problems subject-area scientists bring to us for solution. The Fisher–Neyman–Pearson theory of hypothesis testing was fashioned for a scientific world where experimentation was slow and difficult, producing small data sets

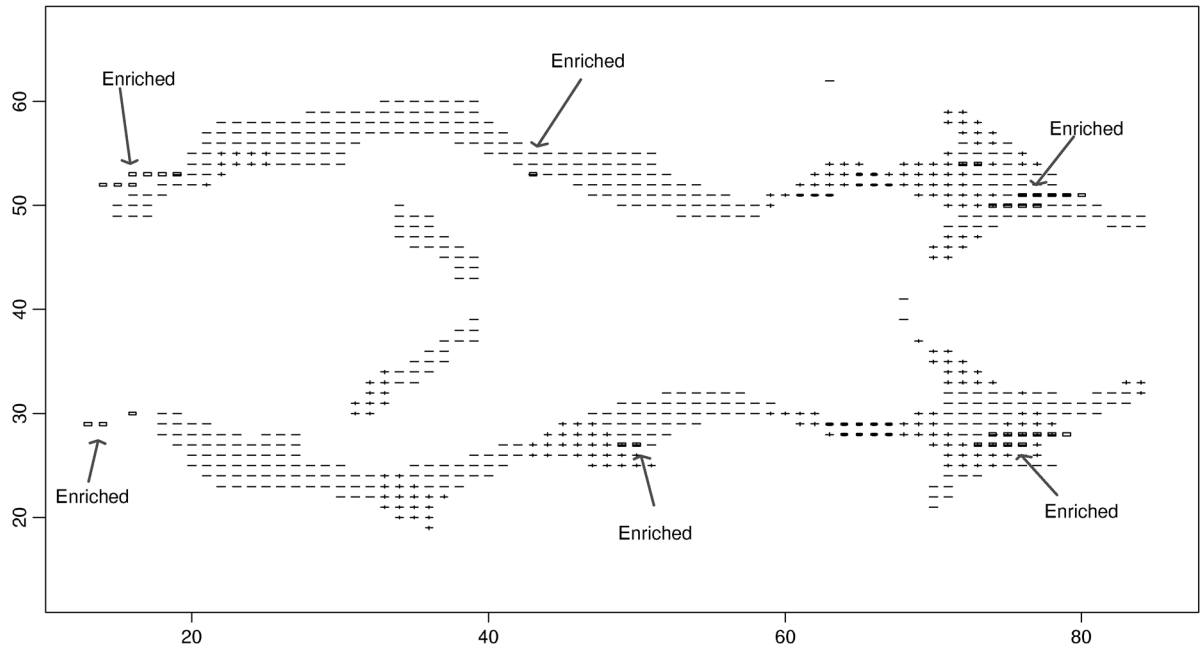

FIG. 10. *Enrichment analysis of Imaging data, panel* D *of Figure* 1; *z-value for original* 15,445 *voxels have been averaged over* "*gene-sets*" *of neighboring voxels with city-block distance* ≤ 2. *Coded as* "$-$" *for* $\bar{z}_i < 0$, "$+$" *for* $\bar{z}_i \geq 0$; *solid rectangles, labeled as* "*Enriched*," *show voxels with* $\widehat{\text{fdr}}(\bar{z}_i) \leq 0.2$, *using empirical null*.

focused on answering single questions. It was wonderfully successful within this milieu, combining elegant mathematics and limited computational equipment to produce dependable answers in a wide variety of application areas.

The three forces have changed relative intensities recently. Computation has become literally millions of times faster and more powerful, while scientific applications now spout data in fire-hose quantities. (Mathematics, of course, is still mathematics.) Statistics is changing in response, as it moves to accommodate massive data sets that aim to answer thousands of questions simultaneously. Hypothesis testing is just one part of the story, but statistical history suggests that it could play a central role: its development in the first third of the twentieth century led directly to confidence intervals, decision theory and the flowering of mathematical statistics.

I believe, or maybe just hope, that our new scientific environment will also inspire a new look at old philosophical questions. Neither Bayesians nor frequentists are immune to the pressures of scientific necessity. Lurking behind the specific methodology of this paper is the broader, still mainly unanswered, question of how one should combine evidence from thousands of parallel but not identical hypothesis testing situations. What I called "empirical Bayes information" accumulates in a way that is not well understood yet, but still has to be acknowledged: in the situations of Figure 4, the frequentist is not free to stick with classical null hypotheses, while the Bayesian cannot use prior (6.13), at least not without the risk of substantial inferential confusion.

Classical statistics developed in a data-poor environment, as Fisher's favorite description, "small-sample theory," suggests. By contrast, modern-day disciplines such as machine learning seem to struggle with the difficulties of too much data. Both problems, too little and too much data, can afflict microarray studies. Massive data sets like those in Figure 1 are misleadingly comforting in their suggestion of great statistical accuracy. As I have tried to show here, the power to detect interesting specific cases, genes, may still be quite low. New methods are needed, perhaps along the lines of "enrichment," as well as a theory of experimental design explicitly fashioned for large-scale testing situations.

One floor up from the philosophical basement lives the untidy family of statistical models. In this paper I have tried to minimize modeling decisions by working directly with z-values. The combination of the two-groups model and false discovery rates applied to the z-value histogram is notably light on assumptions, more so when using an empirical null, which does not even require independence across the columns of X (i.e., across microarrays, a dangerous assumption as shown

in Section 6 of Efron, 2004). There will certainly be situations when modeling inside the X matrix, as in Newton et al. (2004) or Kerr, Martin and Churchill (2000), yields more information than z-value procedures, but I will leave that for others to discuss.

ACKNOWLEDGMENTS

This research was supported in part by the National Science Foundation Grant DMS-00-72360 and by National Institute of Health Grant 8R01 EB002784.

REFERENCES

ALLISON, D., GADBURY, G., HEO, M., FERNANDEZ, J., LEE, C. K., PROLLA, T. and WEINDRUCH, R. (2002). A mixture model approach for the analysis of microarray gene expression data. *Computat. Statist. Data Anal.* **39** 1–20. MR1895555

AUBERT, J., BAR-HEN, A., DAUDIN, J. and ROBIN, S. (2004). Determination of the differentially expressed genes in microarray experiments using local FDR. *BMC Bioinformatics* **5** 125.

BENJAMINI, Y. and HOCHBERG, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B* **57** 289–300. MR1325392

BENJAMINI, Y. and YEKUTIELI, D. (2001). The control of the false discovery rate under dependency. *Ann. Statist.* **29** 1165–1188. MR1869245

BENJAMINI, Y. and YEKUTIELI, D. (2005). False discovery rate-adjusted multiple confidence intervals for selected parameters. *J. Amer. Statist. Assoc.* **100** 71–93. MR2156820

BROBERG, P. (2005). A comparative review of estimates of the proportion unchanged genes and the false discovery rate. *BMC Bioinformatics* **6** 199.

DO, K.-A., MUELLER, P. and TANG, F. (2005). A Bayesian mixture model for differential gene expression. *J. Roy. Statist. Soc. Ser. C* **54** 627–644. MR2137258

DUDOIT, S., SHAFFER, J. and BOLDRICK, J. (2003). Multiple hypothesis testing in microarray experiments. *Statist. Sci.* **18** 71–103. MR1997066

EFRON, B. (2003). Robbins, empirical Bayes, and microarrays. *Ann. Statist.* **31** 366–378. MR1983533

EFRON, B. (2004). Large-scale simultaneous hypothesis testing: The choice of a null hypothesis. *J. Amer. Statist. Assoc.* **99** 96–104. MR2054289

EFRON, B. (2005). Local false discovery rates. Available at http://www-stat.stanford.edu/~brad/papers/False.pdf.

EFRON, B. (2006). Size, power, and false discovery rates. Available at http://www-stat.stanford.edu/~brad/papers/Size.pdf. *Ann. Appl. Statist.* To appear.

EFRON, B. (2007). Correlation and large-scale simultaneous significance testing. *J. Amer. Statist. Assoc.* **102** 93–103. MR2293302

EFRON, B. and GOUS, A. (2001). Scales of evidence for model selection: Fisher versus Jeffreys. *Model Selection IMS Monograph* **38** 208–256. MR2000754

EFRON, B. and MORRIS, C. (1975). Data analysis using Stein's estimator and its generalizations. *J. Amer. Statist. Assoc.* **70** 311–319. MR0391403

EFRON, B. and TIBSHIRANI, R. (1996). Using specially designed exponential families for density estimation. *Ann. Statist.* **24** 2431–2461. MR1425960

EFRON, B. and TIBSHIRANI, R. (2006). On testing the significance of sets of genes. Available at http://www-stat.stanford.edu/~brad/papers/genesetpaper.pdf. *Ann. Appl. Statist.* To appear.

EFRON, B. and TIBSHIRANI, R. (2002). Empirical Bayes methods and false discovery rates for microarrays. *Genetic Epidemiology* **23** 70–86.

EFRON, B., TIBSHIRANI, R., STOREY, J. and TUSHER, V. (2001). Empirical Bayes analysis of a microarray experiment. *J. Amer. Statist. Assoc.* **96** 1151–1160. MR1946571

HEDENFALK, I., DUGGEN, D., CHEN, Y., ET AL. (2001). Gene expression profiles in hereditary breast cancer. *New Engl. J. Medicine* **344** 539–548.

HELLER, G. and QING, J. (2003). A mixture model approach for finding informative genes in microarray studies. Unpublished manuscript.

KERR, M., MARTIN, M. and CHURCHILL, G. (2000). Analysis of variance in microarray data. *J. Comp. Biology* **7** 819–837.

LANGASS, M., LINDQUIST, B. and FERKINSTAD, E. (2005). Estimating the proportion of true null hypotheses, with application to DNA microarray data. *J. Roy. Statist. Soc. Ser. B* **67** 555–572. MR2168204

LEE, M. L. T., KUO, F., WHITMORE, G. and SKLAR, J. (2000). Importance of replication in microarray gene expression studies: Statistical methods and evidence from repetitive cDNA hybridizations. *Proc. Natl. Acad. Sci.* **97** 9834–9838.

LEHMANN, E. and ROMANO, J. (2005). Generalizations of the familywise error rate. *Ann. Statist.* **33** 1138–1154. MR2195631

LEHMANN, E. and ROMANO, J. (2005). *Testing Statistical Hypotheses*, 3rd ed. Springer, New York. MR2135927

LEWIN, A. RICHARDSON, S., MARSHALL, C., GLASER, A. and AITMAN, Y. (2006). Bayesian modeling of differential gene expression. *Biometrics* **62** 1–9. MR2226550

LIANG, C., RICE, J., DE PATER, I., ALCOCK, C., AXELROD, T., WANG, A. and MARSHALL, S. (2004). Statistical methods for detecting stellar occultations by Kuiper belt objects: The Taiwanese-American occultation survey. *Statist. Sci.* **19** 265–274. MR2146947

LIAO, J., LIN, Y., SELVANAYAGAM, Z. and WEICHUNG, J. (2004). A mixture model for estimating the local false discovery rate in DNA microarray analysis. *Bioinformatics* **20** 2694–2701.

NEWTON, M., KENDZIORSKI, C., RICHMOND, C., BLATTNER, F. and TSUI, K. (2001). On differential variability of expression ratios: Improving statistical inference about gene expression changes from microarray data. *J. Comp. Biology* **8** 37–52.

NEWTON, M., NOVEIRY, A., SARKAR, D. and AHLQUIST, P. (2004). Detecting differential gene expression with a semiparametric hierarchical mixture model. *Biostatistics* **5** 155–176.

PAN, W., LIN, J. and LE, C. (2003). A mixture model approach to detecting differentially expressed genes with microarray data. *Functional and Integrative Genomics* **3** 117–124.

PARMIGIANI, G., GARRETT, E., AMBAZHAGAN, R. and GABRIELSON, E. (2002). A statistical framework for expression-based molecular classification in cancer. *J. Roy. Statist. Soc. Ser. B* **64** 717–736. MR1979385

PAWITAN, Y., MURTHY, K., MICHIELS, J. and PLONER, A. (2005). Bias in the estimation of false discovery rate in microarray studies. *Bioinformatics* **21** 3865–3872.

POUNDS, S. and MORRIS, S. (2003). Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of the p-values. *Bioinformatics* **19** 1236–1242.

QUI, X., BROOKS, A., KLEBANOV, L. and YAKOVLEV, A. (2005). The effects of normalization on the correlation structure of microarray data. *BMC Bioinformatics* **6 120**.

ROGOSA, D. (2003). Accuracy of API index and school base report elements: 2003 Academic Performance Index, California Department of Education. Available at http://www.cde.cagov/ta/ac/ap/researchreports.asp.

SCHWARTZMAN, A., DOUGHERTY, R. F. and TAYLOR, J. E. (2005). Cross-subject comparison of principal diffusion direction maps. *Magnetic Resonance in Medicine* **53** 1423–1431.

SINGH, D., FEBBO, P., ROSS, K., JACKSON, D., MANOLA, J., LADD, C. TAMAYO, P., RENSHAW, A., D'AMICO, A., RICHIE, J., LANDER, E., LODA, M., KANTOFF, P., GOLUB, T. and SELLERS, R. (2002). Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell* **1** 302–309.

SMYTH, G. (2004). Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.* **3** 1–29. MR2101454

STOREY, J. (2002). A direct approach to false discovery rates. *J. Roy. Statist. Soc. Ser. B* **64** 479–498. MR1924302

STOREY, J., TAYLOR, J. and SIEGMUND, D. (2005). Strong control conservative point estimation and simultaneous conservative consistency of false discovery rates: A unified approach. *J. Roy. Statist. Soc. Ser. B* **66** 187–205.

SUBRAMANIAN, A., TAMAYO, P., MOOTHA, V. K., MUKHERJEE, S., EBERT, B. L., GILLETTE, M. A., PAULOVICH, A., POMEROY, S. L., GOLUB, T. R., LANDER, E. S. and MESIROV, J. P. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci.* **102** 15545–15550.

TURNBULL, B. (2006). BEST proteomics data. Available at www.stanford.edu/people/brit.turnbull/BESTproteomics.pdf.

TUSHER, V., TIBSHIRANI, R. and CHU, G. (2001). Significance analysis of microarrays applied to transcriptional responses to ionizing radiation. *Proc. Natl. Acad. Sci. USA* **98** 5116–5121.

VAN'T WOUT, A., LEHRMA, G., MIKHEEVA, S., O'KEEFFE, G., KATZE, M., BUMGARNER, R., GEISS, G. and MULLINS, J. (2003). Cellular gene expression upon human immunodeficiency virus type 1 infection of CD$+ T-Cell lines. *J. Virology* **77** 1392–1402.

Statistical Science
2008, Vol. 23, No. 1, 23–28
DOI: 10.1214/07-STS236B
Main article DOI: 10.1214/07-STS236

Comment: Microarrays, Empirical Bayes and the Two-Groups Model

Yoav Benjamini

Efron has given us a comprehensive and thoughtful review of his approach to large-scale testing stemming from the challenges of analyzing microarray data. Addressing the microarray challenge right from the emergence of the technology, and adapting the point of view on multiple testing that emphasizes the false discovery rate, Efron's contributions in both fields have been immense. In the discussed paper he reviews philosophy, motivation, methodologies and even practicalities, and along this process gives us a view of the field of statistics from an eagle's eye.

A thorough discussion of such a work is a major undertaking. Instead, I shall first comment on five issues and then discuss new directions for research on large-scale multiple inference that bear on Efron's review. The scope of the review challenges a discussant to try and address some of the issues raised from a broader point of view. I shall give it a try.

1. FDR AND LOCAL FDR

Efron notes some of the practical difficulties with the local version of FDR that relies on densities: densities are more difficult to estimate, with higher variability and stronger reliance on assumptions about the tails. These difficulties are even more pronounced in the far tails, where the estimation of the null is more problematic, yet this is where they are usually calculated.

Not surprisingly, I prefer to report and control the FDR $= E(V/R)$, using the usual notation of R being the number rejected, V the number falsely rejected and treating the ratio as 0 when $R = 0$. But my reason is not technical, in the sense that working with cumulative distribution function is easier than with densities, but a fundamental one: the concern to assure reproducible results in the face of selection effects. In other words, how is our inference affected by the fact that we only select as discoveries genes passing some (data-dependent) threshold? Will the identification of discoveries be reproducible when another experimenter will look at a similar question using new data?

The local false discovery rate fdr(z) does not cater to the selection process, but rather to the observed value of the statistics and to the abundance of potential discoveries in the pool studied. If it is then used for selection, say by identifying all genes with local FDR values above a threshold, or by displaying all such voxels on a map, the property of the set of genes selected this way is unknown.

Such tail-selection is not only evident in the way results are reported in the examples discussed by Efron, but is needed in order to address the replicability of the results. Identifying a specific gene with a small fdr(z), it is unlikely that the result will be replicated in another study, in the sense that the fdr(z) for that gene will be similar. One can still require that the result in the repeated experiment be at least as extreme as the one in the original experiment, possibly relative to a common standard such as the 0.05.

Both arguments call for the evaluation of tail probability (or expectation), thereby considering the effects of selecting all results that are more extreme than a (data-dependent) threshold. Therefore, even in the empirical Bayes framework, the assessment of the selection effect should be by the Fdr (and pFDR) rather than with the local FDR.

In summary, my position is that for the initial screening of potential genes, for the purpose of creating a pool of candidate hypotheses for further study, and *for reporting the results of an experiment in the literature*, I would strongly argue for the use of tail-based measures such as the FDR. The local FDR may still be useful for the decision-making scientist: the pool of candidate genes is not the end of the story in an investigation, but rather its beginning. More comparisons, literature survey and database searches are commonly done, before, say, a wet lab experiment is conducted on a few genes. When it comes to making personal decisions as to what leads to follow with more extensive research, within the previously identified set, local FDR

Y. Benjamini is Professor, Department of Statistics and Operations Research, Tel Aviv University, Tel Aviv, Israel (e-mail: ybenja@post.tau.ac.il).

about a specific gene can give valuable information together with effects size assessment (see below) and the other relevant data gathered.

2. THE FDR, THE pFDR OR THE Fdr

The difference between the Bayesian and empirical Bayesian approach on the one hand, and the frequentist approach on the other hand, seems to surface in the form of emphasis on pFDR and Fdr (and fdr) rather than on FDR. Yet in the often used mixture model for microarray analysis, with n large and $p_0 < 1$, for a thresholding procedure at a fixed threshold,

$$\begin{aligned}\mathrm{FDR} &= E(V/R)\\ &= E(V/R|R>0)\Pr(R>0)\\ &\approx E(V/R|R>0)\\ &= p\mathrm{FDR}\\ &\approx E(V)/E(R) = \mathrm{Fdr}.\end{aligned}$$

The difference emerges as we deal with problems where $p_0 = 1$ is a real possibility: a trait is not related to any genetic factor, say. In this case the control of FDR is similar to the control of $\Pr(R > 0)$, which in this case is the probability of making any error, while the other concepts are identically 1. So the original FDR, interpretable as a Bayesian concept in one situation, turns out to be the classical frequentist familywise error-rate in another.

The frequentist approach emerges to be useful in yet another important situation, that of estimating sparse signals. As Abramovich et al. (2006) show, in the case when the number of tests grows to infinity, it is optimal in a minimax sense to first use FDR controlling testing and then estimate the significant parameters, thresholding the others to 0. In fact, even when $p_0 = 0$, that is, all parameters are different from 0, but the size of the ordered parameters decays quickly to 0, such FDR-testimation is optimal.

3. ESTIMATING THE PROPORTION OF NULL HYPOTHESES

The first versions of our work on FDR included an estimator of the proportion of null hypotheses p_0. The inability to publish this work, which had but a single simple theorem and many simulations, led us to drop the adaptive stage where p_0 was estimated and replace it by 1, thereby enabling us to get the elegant proof. I believe the original editors did us a favor by requiring mathematical serenity, which led to Benjamini and Hochberg (1995). But they erred when they consequently still refused to publish both our original adaptive results where p_0 was estimated, as well as those of Williams, Jones and Tukey, for too strong a reliance on computer work (both appeared in 2000). I think that the right mixture of the two is needed in statistics, as exemplified in Efron's own work.

In the density mixture model it is usually beneficial to estimate p_0. Elsewhere, it may be more useful to bound the number of extremely small parameters, rather than to estimate the number exactly at 0: it does not make sense to estimate p_0 in the sparsity model from Abramovich et al. (2006), where no parameter is at 0; still it may turn out beneficial to use a bound on the number of parameters nondistinguishable from 0. In the two-stage procedure of Benjamini, Krieger and Yekutieli (2006) such a bound is offered by $n - R(1 - q)$. In the adaptive step-down procedure described there, once rejecting $i - 1$ hypotheses, a new bound on n_0 is offered by $n_{0i} = n + 1 - i(1 - q)$. The procedure steps from $i = 1$ on, as long as $p_{(i)} \le qi/n_{0i}$. The FDR controlling property of this procedure is given in Gavrilov, Benjamini and Sarkar (2008), and its asymptotic optimality follows from Finner, Dickhaus and Roters (2008).

4. EFFECT OF DEPENDENCY

There is a misconception that pFDR and local Fdr do not require independence while FDR does. Quoting Efron, false discovery rate control is verified for the procedure in BH "under the assumption of independence among the N z-values (relaxed a little)," and that this seems fatal for microarray applications. Yet on the other hand, "a great virtue of the empirical Bayes/two-groups approach is that independence is not necessary."

First, as noted before, in the asymptotic mixture model advocated for microarray analysis, $\mathrm{FDR} = p\mathrm{FDR} = \mathrm{Fdr} = E(V)/E(R)$ as n tends to ∞. Obviously the last measure is not sensitive to dependency when a fixed threshold is used, meaning that in this model so will the FDR. Moreover, as noted by Efron, both approaches rely on the same statistic, namely on $\widehat{\mathrm{Fdr}}(z)$, so how can dependency be fatal for one and unnecessary for the other?

Away from the mixture model, the measure $\mathrm{Fdr} = E(V)/E(R)$ is indeed not sensitive to dependence when evaluated at a fixed threshold, unlike FDR or pFDR that summarize the distribution of V/R displayed in Figure 6. But that comes at the expense of destroying the dependence between the number of false

discoveries V and the number of discoveries R within the same experiment, as expectations are taken separately. Furthermore, even if the Fdr itself is not sensitive to dependence, its estimator may very well be, and the properties of procedures that take the maximal number of rejections subject to an estimated false discovery measure less than some threshold are prone to have the difficulties mentioned about the procedure in BH under unusual dependency (as in Benjamini and Yekutieli, 2001).

It was mentioned that the procedure in BH controls the FDR under positive regression dependency structure. Even outside this realm, numerous studies indicate the FDR controlling procedure in BH has a robust behavior under the dependency encountered in practical problems. In a systematic study, using a combination of simulations and analytic results, Reiner-Benaim (2007) showed that for two-sided normally distributed test statistics the FDR is always controlled at the desired level q under a wide collection of correlation structures. In extreme situations the FDR may get somewhat higher than qp_0, though, so adaptive methods with estimated p_0 are somewhat more sensitive to dependency. Interestingly, the structure of constant correlation of, say, all comparisons with same control, where the FDR of the procedure in BH is assured to be less than q, is not covered by the current asymptotic results for the pFDR or local FDR. These results require consistency of the empirical distribution of the p-values as the number of hypotheses tends to infinity.

In summary, the appropriate statement regarding dependence should be much more balanced than the simple statement that FDR has a problem under dependency and Fdr and local FDR do not. Both approaches are quite robust to the dependence structures encountered in microarray studies, and both are more sensitive when estimators of p_0 are incorporated, but not in a critical way.

5. ESTIMATING THE DISTRIBUTION UNDER THE NULL

A central theme in Efron's work is the opportunity that large problems offer for estimating the components of the statistical model that are usually treated as assumed. In particular Efron emphasizes rightfully that in many microarray datasets the distribution of the p-values evaluated under the assumed null distribution is not uniform, as can be seen either directly or from the nonnormality of the z-transformed p-values. They are in fact far from normal even in the center, where they should mostly come from the true null hypotheses. Four possible reasons for the discrepancy are discussed. Motivated by the empirical Bayes approach, the estimation of the distribution under the null is offered as a remedy. This remedy may prove useful for frequentist analysis as well.

I agree to all four sources of problems offered by Efron, but would like to offer a fifth one: the set of p-values reaching the stage of statistical analysis has been selected from the set originally measured. I noticed this phenomenon a while ago, and commented in a highly prestigious genomics conference that this is not an innocent act. I was almost booed, and the impact of the rest of my lecture diminished, I am afraid. But then, take the microarray examples discussed by Efron: for the three microarray datasets I tried to trace back the reason for the particular number of genes reported in the analysis.

For the HIV data example, the Methods section in the original publication explains: "We used a standard deviation threshold of 50 expression units to select the most variable transcript sequences." For the Singh et al. (2002) prostate data, "Genes whose expression varied less than 5-fold between any two samples in any given experiment were removed." In the BRCA dataset the selection is even more severe. There were 5361 unique genes measured; only 3226 are analyzed. The reason: "In the analyses involving cDNA microarrays, a total of 3226 genes with an average intensity (level of expression) of more than 2500 pixels among all samples, an average spot area of more than 40 pixels, and no more than one sample in which the size of the spot area was 0 pixels were included."

In the first two cases the effect of the selection is clearly to omit more genes with no differences between the groups. Even in the third case, the selection was not (the possibly legitimate) conditioning on the average of normal variates before testing their differences, but a more complicated selection on the original and nonnormal scale.

I suspect that the above reason is as important as the other four proposed for the discrepancy between the real distribution and the assumed one. Worse than that, it affects the center of the distribution under the null usually in a way that distorts the connection between the center and the tails. Therefore inferring about the distribution under the null at the tails from the central part of the empirical distribution is very precarious. For examples of such effects, notice the dips at the center of the actual distributions relative to the estimated distributions under the null in Figure 4(a), and relative to the theoretical null in Figure 1(a).

I do not have a full answer to this difficulty. With some datasets we found that careful preprocessing solved much of the problem. In an extremely large and complicated problem we still struggle, starting from all measured expression data. My point is that practically, certainly with microarray data, I am still more comfortable using an appropriately verified theoretical null distribution than an estimated one. I shall be more confident about estimation if it is tailored to handle the effects of the selection process that hides behind the regular technical microarray preprocessing analysis, a step usually masked from the statisticians' eyes.

With this I end my comments about some of the methodologies offered and opinions expressed, and turn to comment on the future of multiple hypotheses testing in large genomic problems.

6. CONFIDENCE INTERVALS FOLLOWING SELECTION

This issue is rarely addressed in the large significance screening studies, so I am happy to find it emphasized in Efron's paper. Too often the decisions as to what clues to follow with expensive research are based on significance only, rather than on estimated effect sizes. The latter calls for making confidence statements about the few selected parameters.

I do not necessarily find here a possible clash between the frequentist and the empirical Bayes approaches. The optimality result in Benjamini and Yekutieli (2005), which is being viewed as evidence that the two approaches are on clashing orbits, is stated only for two-sided symmetric and equivariant confidence intervals (having the same shape under translation and reflection). Frequentist confidence intervals need not be equivariant. Allowing such flexibility, a confidence statement with a special role for 0 is not a result of Bayesian analysis only: a confidence interval that includes both 0 and an interval not connected to 0 may emerge as a result of inverting nonequivariant acceptance regions, as shown and discussed in Benjamini, Hochberg and Stark (1998).

Interestingly, in his recent talk in MCP2007, Yekutieli presented a case where Bayesian intervals constructed to incorporate the selection effect used, enjoy False Coverage Rate properties. This indicates that there is potential benefit for successful research on setting confidence intervals after selection, and that pursuing this goal from all approaches, frequentist, Bayes, and empirical Bayes, hold better promise for rapid developments, as was the case for testing.

7. THE TRANSITION FROM VERY LARGE TO HUGE PROBLEMS

I cannot agree more with Efron when he states that applications is one of the three fundamental forces influencing statistics. Our motivation in developing FDR has been the analysis of clinical trials in which there were 100 or so endpoints. The FDR criterion turned out to be inherently scalable, in the sense that it has stood up to the challenges of tens of thousands of hypotheses. It is becoming more common now to search over millions of hypotheses looking for discoveries against the noisy background (see below), and the FDR is still relevant—possibly because of its triple Frequentist/ Empirical Bayes/Bayes interpretation.

Still, the tools developed along with the approach may have reached the stage where it is unlikely that further polishing of same tools will be of much help: for example, better estimation of p_0 will offer little improvement, because in such huge problems p_0 is very close to 1 or even 1; the discoveries are not likely to be important if abundant in the extremely large pool searched. That the tools are polished is in fact a tribute to the many researchers in the statistical community who in their (sometimes competitive) efforts advanced tremendously our knowledge and understanding about false discovery rates.

There are three possible directions to deal with the new challenges of huge multiple testing problems using new tools, and almost all are related in some sense to the approach presented by Efron.

D1. The enrichment analysis offered by Efron points at an important first direction: increasing signal to noise ratio by collecting hypotheses to sets in which they are likely to be true together, or false together. In Efron's gene-enrichment example the clustering of genes into sets is based on external information regarding the pathways involved. In the brain-imaging example the clusters are based on a moving window (for FDR controlling scan statistics see also Pacifico et al., 2007). The cluster-based analysis in Benjamini and Heller (2007) is another example of the enrichment approach in the brain-imaging problem. It comes in a different flavor, though: based on the pilot study routinely performed in brain-imaging experiments, the voxels in the brain are first clustered to create a coarser partition of the hypotheses. The clusters need not be of the same size and shape, and are of neurological relevance. Then, in the main experiment, the clusters are tested using a combining statistic for each cluster. Therefore, not only do we gain increased power from

combining the evidence over clusters, but also when addressing multiplicity the number of clusters tested is much smaller than the number of voxels tested.

The essence of the above examples is clear. When the tested parameters have further structure, in the sense that we can have a grasp in what sets the hypotheses are going to be true together and false together (correlated parameters), enrichment analysis is of great potential: in many cases not only is the signal to noise ratio increased but the multiplicity problem can be reduced.

D2. The second direction is that of employing weights to differentiate between the hypotheses tested. The weights may incorporate differing importance of the hypotheses (Benjamini and Hochberg, 1997), or different prospects for showing effects (Genovese, Roeder and Wasserman, 2006). As in the case of enrichment analysis, the weights can be based on outside information, or on information from initial testing. Weights can also answer Efron's third concern regarding the enrichment analysis, especially as one may also assign weights to sets of hypotheses in a way that is proportional to the size of the set, as is done in Benjamini and Heller (2007).

D3. The third direction is that of endowing a hierarchical structure to the family of hypotheses tested, where a subfamily of hypotheses at a branch is tested only after the node from which it branches is tested and rejected. When such hierarchical structure enjoys an enrichment property, again in the sense that the hypotheses in a branch tend to be true or false together, an opportunity for power gain arises. Reiner et al. (2007), for example, test the association between the expression of some 27K genes in each of five brain regions, and 17 behavioral traits, a study of more than 2.2 million hypotheses. They first screen for genes and brain regions where strain differences exist. Only those combinations of brain regions and genes passing an FDR-based threshold are further tested for correlation, each subfamily of 17 tests of correlation being tested separately. The theoretical questions regarding such procedures are discussed in Yekutieli (2008), and answers are given there for the case where the test statistics at a node and at its branching hypotheses are independent. In Benjamini and Heller (2007) described above, a natural hierarchy is to test clusters of voxels and then individual voxels within clusters. The test statistics for testing a cluster and those for testing voxels within a rejected cluster are not independent, so conditional p-values have to be estimated for their hierarchical use.

Recent work by Meinhausen (2008) for testing the importance of a large number of variables in a regression model also makes use of a hierarchical approach, this time within the familywise error-rate framework. As to the design questions for very large experiments raised by Efron, they can also be answered within a hierarchical setting. Zehetmayer, Bauer and Posch (2005), for example, first use a screening experiment, where each hypothesis is tested with no attention to multiplicity, and based on new data conditional p-values for the hypotheses that pass the initial screening are calculated and multiplicity is addressed by FDR control.

Most of these efforts seem to indicate that managing to create hierarchical testing structures that enjoy enrichment properties is extremely promising. Tools such as the estimated distribution under the null, the estimated p_0 at a branch, and sometimes the local FDR, which emerged from the empirical Bayes perspective, can remain useful but may need further adjustments. Allow me to remain philosophical, at this point, and not dwell on details (especially since I do not know how to handle them).

8. IN CONCLUSION

I enjoyed reading the first sections offering a broad view of past achievements, even when I disagree with some of the solutions offered. I am enthusiastic about the last three sections, in which new directions of progress are identified to answer needs in applications, as we pass from very large problems to huge ones. I found it important to emphasize the more general nature of these directions, and connect them to current research efforts that reflect similar attitudes. I thank again Efron for taking such a broad view of the subject, thereby calling for a more-than-technical discussion on my part.

ACKNOWLEDGMENTS

I would like to thank Dani Yekutieli for engaging discussions about the above ideas. This work was supported by grants from GIF, NIH and ISF.

REFERENCES

ABRAMOVICH, F., BENJAMINI, Y., DONOHO, D. and JOHNSTONE, I. M. (2006). Adapting to unknown sparsity by controlling the false discovery rate. *Ann. Statist.* **34** 584–653. MR2281879

BENJAMINI, Y. and HELLER, R. (2007). False discovery rate for spatial data. *J. Amer. Statist. Assoc.* **102** 1272–1281.

BENJAMINI, Y. and HOCHBERG, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B* **57** 289–300. MR1325392

BENJAMINI, Y. and HOCHBERG, Y. (1997). Multiple hypotheses testing with weights. *Scand. J. Statist.* **24** 407–418. MR1481424

BENJAMINI, Y., HOCHBERG, Y. and STARK, P. B. (1998). Confidence intervals with more power to determine the sign: Two ends constrain the means. *J. Amer. Statist. Assoc.* **93** 309–317. MR1614569

BENJAMINI, Y., KRIEGER, A. M. and YEKUTIELI, D. (2006). Adaptive linear step-up procedures that control the false discovery rate. *Biometrika* **93** 491–507. MR2261438

BENJAMINI, Y. and YEKUTIELI, D. (2001). The control of the false discovery rate under dependency. *Ann. Statist.* **29** 1165–1188. MR1869245

BENJAMINI, Y. and YEKUTIELI, D. (2005). False discovery rate-adjusted multiple confidence intervals for selected parameters. *J. Amer. Statist. Assoc.* **100** 71–93. MR2156820

FINNER, H., DICKHAUS, T. and ROTERS, M. (2008). On the false discovery rate and asymptotically optimal rejection curve. *Ann. Statist.* To appear.

GAVRILOV, Y., BENJAMINI, Y. and SARKAR, S. (2008). An adaptive step-down procedure with proven FDR control. *Ann. Statist.* To appear.

GENOVESE, C. R., ROEDER, K. and WASSERMAN, L. (2006). False discovery control with p-value weighting. *Biometrika* **93** 509–524. MR2261439

MEINSHAUSEN, N. (2008). Hierarchical testing of variable importance. *Biometrika*. To appear.

PACIFICO, M. P., GENOVESE, C., VERDINELLI, I. and WASSERMAN, L. (2007). Scan clustering: A false discovery approach. *J. Multivariate Anal.* **98** 1441–1469. MR2364129

REINER-BENAIM, A. (2007). FDR control by the BH procedure for two sided correlated tests with implications to gene expression data analysis. *Biometrical J.* **49** 107–126.

REINER-BENAIM, A., YEKUTIELI, D., LETWIN, N. E., ELMER, G. I., LEE, N. H., KAFKAFI, N. and BENJAMINI, Y. (2007). Associating quantitative behavioral traits with gene expression in the brain: Searching for diamonds in the hay. *Bioinformatics* **23** 2239–2246.

SINGH, D., FEBBO, P., ROSS, K., JACKSON, D., MANOLA, J., LADD, C., TAMAYO, P., RENSHAW, A., D'AMICO, A., RICHIE, J., LANDER, E., LODA, M., KANTOFF, P., GOLUB, T. and SELLERS, R. (2002). Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell* **1** 302–309.

YEKUTIELI, D. (2008). Hierarchical false discovery rate controlling methodology. *J. Amer. Statist. Assoc.* To appear.

ZEHETMAYER, S., BAUER, P. and POSCH, M. (2005). Two-stage designs for experiments with a large number of hypotheses. *Bioinformatics* **21** 3771–3777.

Statistical Science
2008, Vol. 23, No. 1, 29–33
DOI: 10.1214/07-STS236C
Main article DOI: 10.1214/07-STS236

Comment: Microarrays, Empirical Bayes and the Two-Group Model

T. Tony Cai

Professor Efron is to be congratulated for his innovative and valuable contributions to large-scale multiple testing. He has given us a very interesting article with much material for thought and exploration. The two-group mixture model (2.1) provides a convenient and effective framework for multiple testing. The empirical Bayes approach leads naturally to the local false discovery rate (Lfdr) and gives the Lfdr a useful Bayesian interpretation. This and other recent papers of Efron raised several important issues in multiple testing such as theoretical null versus empirical null and the effects of correlation. Much research is needed to better understand these issues.

Virtually all FDR controlling procedures in the literature are based on thresholding the ranked p-values. The difference among these methods is in the choice of the threshold. In multiple testing, typically one first uses a p-value based method such as the Benjamini–Hochberg procedure for global FDR control and then uses the Lfdr as a measure of significance for individual nonnull cases. See, for example, Efron (2004, 2005). In what follows I will first discuss the drawbacks of using p-value in large-scale multiple testing and demonstrate the fundamental role played by the Lfdr. I then discuss estimation of the null distribution and the proportion of the nonnulls. I will end with some comments about dealing with the dependency.

In the discussion I shall use the notation given in Table 1 to summarize the outcomes of a multiple testing procedure.

With the notation given in the table, the false discovery rate (FDR) is then defined as $\mathrm{FDR} = \mathrm{E}(N_{10}/R|R > 0)\Pr(R > 0)$.

1. THE USE OF *p*-VALUES: VALIDITY VERSUS EFFICIENCY

In the classical theory of hypothesis testing the p-value is a fundamental quantity. For example, the decision of a test can be made by comparing the p-value with the prespecified significance level α. In the more recent large-scale multiple testing literature, p-value continues to play a central role. As mentioned earlier, nearly all FDR controlling procedures separate the nonnull hypotheses from the nulls by thresholding the ordered p-values.

A dual quantity to the false discovery rate is the false nondiscovery rate $\mathrm{FNR} = \mathrm{E}(N_{01}/S|S > 0)\Pr(S > 0)$. In a decision-theoretical framework, a natural goal in multiple testing is to find, among all tests which control the FDR at a given level, the one which has the smallest FNR. We shall call an FDR procedure *valid* if it controls the FDR at a prespecified level α, and *efficient* if it has the smallest FNR among all FDR procedures at level α. The literature on FDR controlling procedures so far has focused virtually exclusively on the validity; the efficiency issue has been mostly untouched.

In a recent article, Sun and Cai (2007) considered the multiple testing problem from a compound decision point of view. It is demonstrated that p-value is in fact not a fundamental quantity in large-scale multiple testing; the local false discovery rate (Lfdr) is. Thresholding the ordered p-values does not in general lead to efficient multiple testing procedures. The reason for the inefficiency of the p-value methods can be traced back to Copas (1974) where a weighted classification problem was considered. Copas (1974) showed that if a symmetric classification rule for dichotomies is admissible, then it must be ordered by the likelihood ratios, which is equivalent to being ordered by the Lfdr. Sun and Cai (2007) showed that, under mild conditions, the multiple testing problem is in fact equivalent to the weighted classification problem. I will discuss below some of the findings in Sun and Cai (2007) and draw connections to the present paper by Professor Efron.

The local false discovery rate, defined in (2.7), was first introduced in Efron et al. (2001) as the a posteriori probability of a gene being in the null group given the z-score z. The results in Sun and Cai (2007) show that the Lfdr is a fundamental quantity which can be used directly for optimal FDR control. By using the Lfdr

T. Tony Cai is Dorothy Silberberg Professor of Statistics, Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA (e-mail: tcai@wharton.upenn.edu).

TABLE 1

	Claimed nonsignificant	Claimed significant	Total
Null	N_{00}	N_{10}	m_0
Nonnull	N_{01}	N_{11}	m_1
Total	S	R	m

directly for testing, the goals of global error control and individual case interpretation are naturally unified.

For convenience, in the following we shall work with the marginal false discovery rate $\mathrm{mFDR} = \mathrm{E}(N_{10})/\mathrm{E}(R)$ and the marginal false nondiscovery rate $\mathrm{mFNR} = \mathrm{E}(N_{01})/\mathrm{E}(S)$. The mFDR is asymptotically equivalent to the usual FDR under weak conditions, $\mathrm{mFDR} = \mathrm{FDR} + O(m^{-1/2})$, where m is the number of hypotheses. See Genovese and Wasserman (2002).

It is illustrative to first look at an example in the so-called oracle setting, where in the two-group mixture model (2.6) the proportion p_0, the density f_0 of the null distribution and the density f of the marginal distribution are assumed to be known. In this case, both the optimal threshold for the p-values and the optimal threshold for the Lfdr values can be calculated for any given mFDR level. We shall call a testing procedure with the optimal cutoff the *oracle procedure*. Suppose the z-values $z_1, \ldots, z_m$ come from a normal mixture distribution with

$$f(z) = p_0\phi(z) + p_1\phi(z-\mu_1) + p_2\phi(z-\mu_2), \tag{1}$$

where $p_0 = 0.8$, $p_1 + p_2 = 0.2$. That is, in the two-group model (2.6), the null distribution is $N(0,1)$, the distribution of the nonnulls is a two-component normal mixture, and the total proportion of the nonnulls is 0.2. Figure 1 compares the performance of the p-value and Lfdr oracle procedures (see Sun and Cai, 2007).

In Figure 1, panel (a) plots the mFNR of the two oracle procedures as a function of p_1 in (1) where the mFDR level is set at 0.10, and the means under the alternative are $\mu_1 = -3$ and $\mu_2 = 3$. Panel (b) plots the mFNR as a function of p_1 in the same setting except that the alternative means are $\mu_1 = -3$ and $\mu_2 = 6$. In panel (c) we choose mFDR$= 0.10$, $p_1 = 0.18$, $p_2 = 0.02$, $\mu_1 = -3$ and plot the mFNR as a function of μ_2. Panel (d) plots the mFNR as a

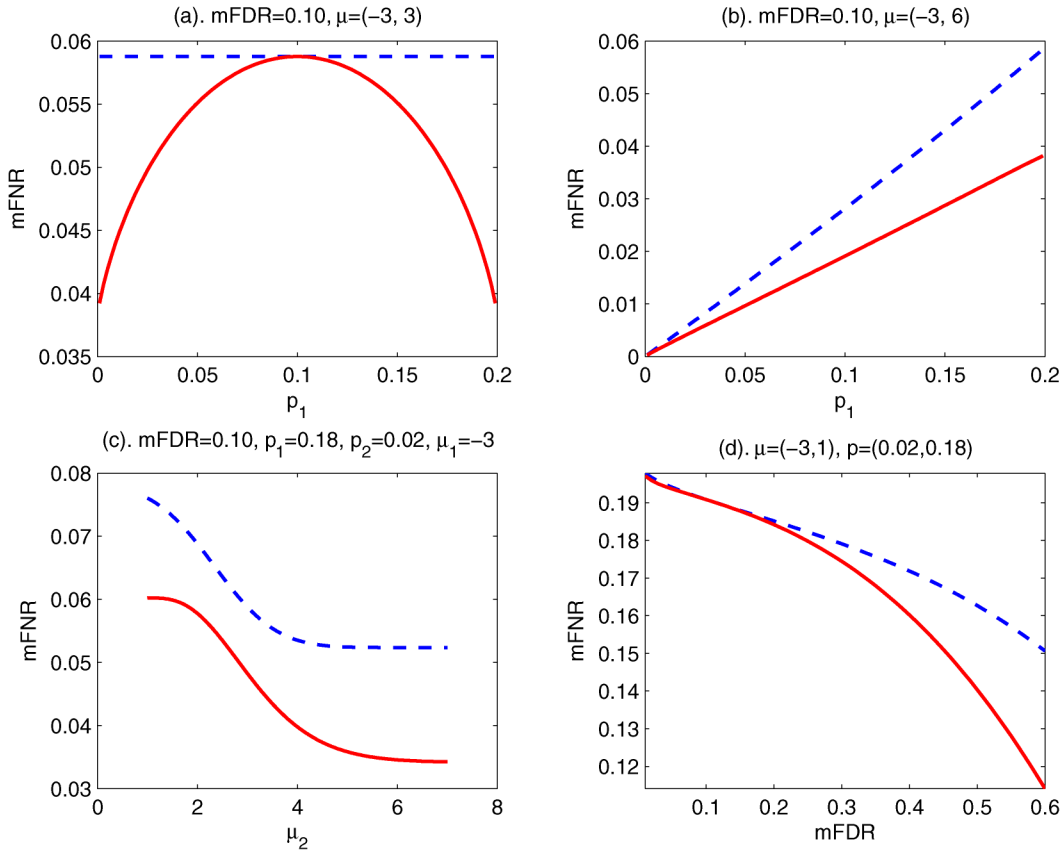

FIG. 1. *The comparison of the p-value (dashed line) and z-value (solid line) oracle rules.*

function of the mFDR level while holding $\mu_1 = -3$, $\mu_2 = 1$, $p_1 = 0.02$, $p_2 = 0.18$ fixed.

It is clear from the plots that the p-value oracle procedure is dominated by the Lfdr oracle procedure. At the same mFDR level, the mFNR of the Lfdr oracle procedure is uniformly smaller than the mFNR of the p-value oracle procedure. The largest difference occurs when $|\mu_1| < \mu_2$ and $p_1 > p_2$, where the alternative distribution is highly asymmetric about the null. When $|\mu_1| = |\mu_2|$, the mFNR remains a constant for the p-value oracle procedure, while the mFNR for the Lfdr oracle procedure can be noticeably smaller when p_1 and p_2 are significantly different, in which case the nonnull distribution has a high degree of asymmetry. The Lfdr oracle procedure utilizes the distributional information of the nonnulls, but the p-value oracle procedure does not.

The Lfdr oracle procedure ranks the relative importance of the test statistics according to their likelihood ratios. An interesting consequence of using the Lfdr statistic in multiple testing is that an observation located farther from the null (i.e., a larger absolute z-value or equivalently a smaller p-value) may have a lower significance level. It is therefore possible that the test accepts a more "extreme" observation while rejecting a less extreme observation, which implies that the rejection region is asymmetric. This is not possible for a testing procedure based on the individual p-values, whose rejection region is always symmetric about the null. This can be seen from Figure 2. The left panel compares the mFNR of the p-value oracle procedure and Lfdr oracle procedure and the right panel compares the rejection region in the case of $p_1 = 0.15$. In this case the Lfdr procedure rejects a z-value of -2 (Lfdr $= 0.227$, p-value $= 0.046$) but not a z-value of 3 (Lfdr $= 0.543$, p-value $= 0.003$). More numerical results are given in Sun and Cai (2007). The results show that the Lfdr oracle procedure dominates the p-value procedure in all configurations of the nonnull hypotheses.

The difference between the two procedures can be even more striking when the alternative distribution f_1 is highly concentrated. In this setting, it is possible that the extreme p-values near both 0 and 1 actually all come from the null distribution instead of the nonnull distribution! In such a case, thresholding the p-values

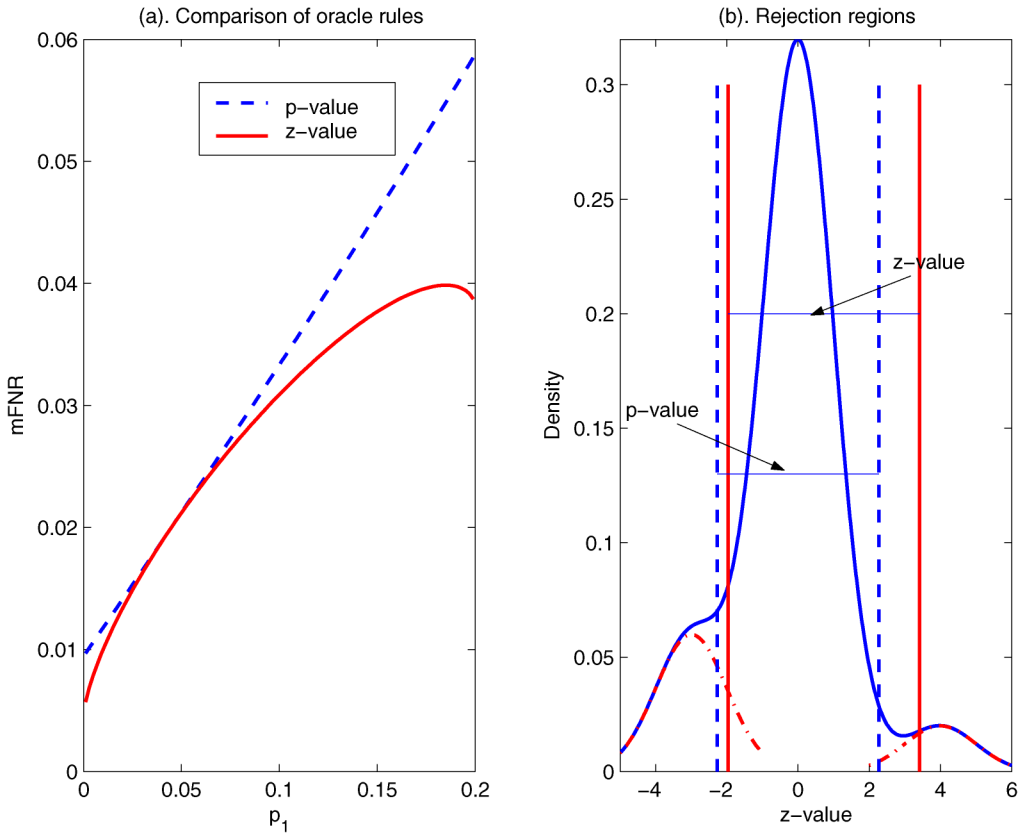

FIG. 2. *Symmetric rejection region versus asymmetric rejection region. In the mixture model* (1), $\mu_1 = -3$ *and* $\mu_2 = 4$. *Both procedures control the mFDR at* 0.10.

fails completely as a method for separating the nonnull hypotheses from the nulls. In contrast, the Lfdr can still be effective in distinguishing between the null and nonnull cases.

In real applications, the proportion p_0 and the density of the marginal distribution f are unknown. With a large number of observed z-values, both p_0 and f can be estimated well from the data. In this regard, the large-scale nature of the problem is a blessing. The null distribution is more subtle. If all the mathematical assumptions are satisfied, the theoretical null distribution is true and thus can be used to compute the Lfdr values. Otherwise, as argued convincingly by Efron in Section 5 of the present paper, the empirical null distribution should be used and it can be estimated from the data. Among the three quantities, p_0, f_0 and f, the marginal density f is relatively easier to estimate than p_0 and f_0. Optimal estimation of these quantities is a challenging problem. We shall discuss the estimation issue in the next section. Let us assume for the moment that we already have consistent estimators $\hat{p}_0$, $\hat{f}_0$ and $\hat{f}$. Such consistent estimators are provided, for example, in Jin and Cai (2007). Define the estimated Lfdr by $\widehat{\text{Lfdr}}(z_i) = [\hat{p}_0 \hat{f}_0(z_i)/\hat{f}(z_i)] \wedge 1$. Sun and Cai (2007) introduced the following adaptive step-up procedure:

$$\text{Let } k = \max\left\{ i : \frac{1}{i}\sum_{j=1}^{i} \widehat{\text{Lfdr}}_{(j)} \le \alpha \right\}, \quad \text{then reject all } H_{(i)},\ i = 1, \ldots, k. \tag{2}$$

It was shown that the data-driven procedure (2) controls the mFDR at level α asymptotically and the mFNR level of the adaptive procedure (2) is asymptotically equal to the mFNR level achieved by the Lfdr oracle procedure. In this sense, the adaptive procedure (2) is asymptotically efficient. Numerical studies in Sun and Cai (2007) show that this adaptive procedure outperforms the step-up procedure (Benjamini and Hochberg, 1995) and the adaptive p-value based procedure (Benjamini and Hochberg, 2000; Genovese and Wasserman, 2004). The numerical results are consistent with the theoretical arguments. These results together show that the Lfdr, not the p-value, is a fundamental quantity for large-scale multiple testing.

It is clear that the performance of the adaptive testing procedure (2) depends to a certain extent on the estimation accuracy of the estimators $\hat{p}_0$, $\hat{f}_0$ and $\hat{f}$. This leads to the estimation issue, which will be discussed next.

2. ESTIMATING THE NULL DISTRIBUTION AND THE PROPORTION OF THE NONNULLS

As demonstrated convincingly in this and other recent papers of Efron, the true null distribution of the test statistic can be quite different from the theoretical null and two seemingly close choices of the null distribution can lead to substantially different testing results. This demonstrates that the problem of estimating the null density f_0 is important to simultaneous multiple testing. In addition to the null density f_0, the proportion of the nonnulls is another important quantity.

Conventional methods for estimating the null parameters are based on either moments or extreme observations. In the present paper, two methods, analytical and geometric, for estimating the null density are discussed. In addition, Efron (2004) suggested an approach which uses the center and half width of the central peak of the histogram for estimating the parameters of the null distribution. These methods are convenient to use. However, the properties of these estimators are still mostly unknown. For example, the analytical method appears to be quite sensitive to the choice of the interval $[a, b]$. It is interesting to understand how the choice of $[a, b]$ affects the resulting estimator $\hat{f}_0$, and more importantly the outcomes of the subsequent testing procedures.

The three null density estimation methods mentioned above rely heavily on the sparsity assumption which means that the proportion of nonnulls is small and most of the z-values near zero come from the nulls. In the nonsparse case these methods of estimating the null densities do not perform well and it is not hard to show that the estimators are generally inconsistent.

Jin and Cai (2007) introduced an alternative frequency domain approach for estimating the null parameters by using the empirical characteristic function and Fourier analysis. The approach demonstrates that the information about the null is well preserved in the high-frequency Fourier coefficients, where the distortion of the nonnull effects is asymptotically negligible. The approach integrates the strength of several factors, including sparsity and heteroscedasticity, and provides good estimates of the null in a much broader range of situations than existing approaches do. The resulting estimators are shown to be uniformly consistent over a wide class of parameters and outperform existing methods in simulations. The approach of Jin and Cai (2007) also yields a uniformly consistent estimator for the proportion of nonnull effects. In a two-component

normal mixture setting, Cai, Jin and Low (2007) proposed an estimator of the proportion and developed a minimax theory for the estimation problem.

Much research is still needed in this area. In particular, it is of significant interest to understand how well the null density can be estimated and how the performance of the estimators affects the performance of the subsequent multiple testing procedures.

3. MODELING THE DEPENDENCY

This paper also raised the important issue of the effects of correlation on outcomes of the testing procedures. Observations arising from large-scale multiple comparison problems are often dependent. For example, different genes may cluster into groups along biological pathways and exhibit high correlation in microarray experiments. It is noted in this paper that correlation can considerably widen or narrow the null distribution of the z-values, and so must be accounted for in deciding which hypotheses should be reported as nonnull. In fact, the notion of null distribution itself becomes unclear in the dependent case.

The focus of previous research on the effects of correlation has been exclusively on the validity of various multiple testing procedures under dependency. For example, Benjamini and Yekutieli (2001) and Wu (2008) showed that the FDR is controlled at the nominal level by the step-up procedure (Benjamini and Hochberg, 1995) and the adaptive p-value procedure (Benjamini and Hochberg, 2000; Storey, 2002; Genovese and Wasserman, 2004) under different dependency assumptions. While the validity issue is important, the efficiency issue is arguably more important.

Intuitively it is clear that the dependency structure among hypotheses is highly informative in simultaneous inference and can be exploited to construct more efficient tests. For example, in comparative microarray experiments, it is found that changes in expression for genes can be the consequence of regional duplications or deletions, and significant genes tend to appear in clusters. Therefore, when deciding the significance level of a particular gene, the observations from its neighborhood should also be taken into account. It is still an open problem how to accommodate the correlation for the construction of valid and efficient multiple testing procedures.

4. CONCLUDING REMARKS

The two-group mixture model and the empirical Bayes approach together provide a useful general framework for multiple testing. The Lfdr, not the p-value, is a fundamental quantity for large-scale multiple testing. The problem of estimating the null density and the proportion of the nonnulls is important to simultaneous multiple testing. This paper raises many important questions and will definitely stimulate new research in the future. I thank Professor Efron for his clear and imaginative work.

ACKNOWLEDGMENT

Research supported in part by NSF Grant DMS-06-04954.

REFERENCES

BENJAMINI, Y. and HOCHBERG, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B* **57** 289–300. MR1325392

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

BENJAMINI, Y. and YEKUTIELI, D. (2001). The control of the false discovery rate under dependency. *Ann. Statist.* **29** 1165–1188. MR1869245

CAI, T., JIN, J. and LOW, M. (2007). Estimation and confidence sets for sparse normal mixtures. *Ann. Statist.* **35** 2421–2449.

COPAS, J. (1974). On symmetric compound decision rules for dichotomies. *Ann. Statist.* **2** 199–204.

EFRON, B. (2004). Large-scale simultaneous hypothesis testing: The choice of a null hypothesis. *J. Amer. Statist. Assoc.* **99** 96–104. MR2054289

EFRON, B. (2005). Local false discovery rates. Available at http://www-stat.stanford.edu/~brad/papers/False.pdf.

EFRON, B., TIBSHIRANI, R., STOREY, J. and TUSHER, V. (2001). Empirical Bayes analysis of a microarray experiment. *J. Amer. Statist. Assoc.* **96** 1151–1160. MR1946571

GENOVESE, C. and WASSERMAN, L. (2002). Operating characteristic and extensions of the false discovery rate procedure. *J. Roy. Statist. Soc. Ser. B* **64** 499–517.

GENOVESE, C. and WASSERMAN, L. (2004). A stochastic process approach to false discovery control. *Ann. Statist.* **32** 1035–1061. MR2065197

JIN, J. and CAI, T. (2007). Estimating the null and the proportion of non-null effects in large-scale multiple comparisons. *J. Amer. Statist. Assoc.* **102** 495–506. MR2325113

STOREY, J. (2002). A direct approach to false discovery rates. *J. Roy. Statist. Soc. Ser. B* **64** 479–498. MR1924302

SUN, W. and CAI, T. (2007). The oracle and compound decision rules for false discovery rate control. *J. Amer. Statist. Assoc.* **102** 901–912.

WU, W. (2008). On false discovery control under dependence. *Ann. Statist.* **36** 364–380.

Statistical Science
2008, Vol. 23, No. 1, 34–40
DOI: 10.1214/08-STS236D
Main article DOI: 10.1214/07-STS236

Comment: Microarrays, Empirical Bayes and the Two-Groups Model

Carl N. Morris

Abstract. Brad Efron's paper has inspired a return to the ideas behind Bayes, frequency and empirical Bayes. The latter preferably would not be limited to exchangeable models for the data and hyperparameters. Parallels are revealed between microarray analyses and profiling of hospitals, with advances suggesting more decision modeling for gene identification also. Then good multilevel and empirical Bayes models for random effects should be sought when regression toward the mean is anticipated.

Key words and phrases: Bayes, frequency, interval estimation, exchangeable, general model, random effects.

1. FREQUENCY, BAYES, EMPIRICAL BAYES AND A GENERAL MODEL

Brad Efron's two-groups approach and the empirical null ("null" refers to a distribution, not to a hypothesis) extension of his local fdr addresses testing many hypotheses simultaneously, with modeling enabled by the repeated presence of many similar problems. He assumes two-level models for random effects, developing theory by drawing on and combining ideas from frequency, Bayesian and empirical Bayesian perspectives. The last half-century in statistics has seen exciting developments from many perspectives for simultaneous estimation of random effects, but there has been little explicit parallel work on the complementary problem of hypothesis testing. That changes in Brad's paper, especially for testing many hypotheses when exchangeability restrictions are plausible.

"Empirical Bayes" is in the paper's title, said in Section 3 to be a "bipolar" methodology that draws on frequency and Bayes, but otherwise with a meaning left for us to infer from the paper's example datasets. The examples all involve two-level models with inferences about many unknown parameters, that is, about the unknown random effects. Blending frequency and Bayesian thinking in statistics will be appreciated especially by statisticians who engage both in theoretical and in applied research, and we know that many of statistics' best and time-honored procedures perform well simultaneously from the frequency and the Bayesian perspectives. Classifying statisticians as either Bayesian or frequentist ignores the fact that these terms have varying meanings to different statisticians, and it encourages the view that statisticians must adopt just one of these perspectives exclusively, which many statisticians, myself included, do not do.

The frequency perspective requires comparing procedures on the basis of repeated sampling, but it can be neutral about how procedures are constructed. The Bayesian approach, after a model is completely specified, including the "prior" ("structural" or "mixing" might be better adjectives) distribution, must use the laws of probability to assess uncertainties about unknowns, given the observed data and the model. Valuably even from the frequency perspective, Bayesian reasoning can be used to suggest how to construct inferences about population parameters and other unobservables, at least in ideal settings. That is illustrated in Efron's treatment of the fdr and the Fdr here. Such modeling of likelihood functions at more than one level, and of priors, becomes less subjective when one has more data, especially with massive datasets like those in the paper.

Scientists have been encountering ever more massive datasets, especially since modern censoring technology and computers have evolved to make collecting, organizing, visualizing and analyzing such databases possible. My early experience in the 1980s involving two-level models for NASA's satellite imagery data made it

Carl N. Morris is Professor, Department of Statistics, Harvard University, Cambridge, Massachusetts 02138, USA (e-mail: morris@stat.harvard.edu)

clear to me that science had reached a new point where computers not only had enabled us to analyze large datasets, they even made it possible to collect very large datasets. The computer had become a horse that could collect and analyze data as we directed, and statisticians were its jockeys. While the massive datasets we see today can be overwhelming, Brad rightly recognizes that they can be welcomed as opportunities to build better models. That not only leads to more accurate inferences for the given data, but better models also advance knowledge and future scientific discoveries.

Brad's use of "empirical Bayes" with the six datasets of the paper is restricted to datasets he considers to be exchangeable. That could signify his moving away from a liberalizing view of empirical Bayes that we once developed together. I doubt this, but the analyses shown assume that the joint distribution of the data and of the random effects are exchangeable. Our papers together in the 1970s moved empirical Bayes away from that requirement, partly to provide a perspective from which acceptable shrinkage generalizations of Stein's estimator might be developed. That was and is needed especially when (nearly) unbiased estimates of the different random effects have different variances, perhaps most often because of different sample sizes.

In seeking a firmer basis for modeling and inference in empirical Bayes settings (Morris, 1983), I continued back then to use that term. However, Herb Robbins, who coined the term, made it clear to me then that the version he had pioneered, built around exchangeability, asymptotics and nonparametric mixing (prior) distributions, was how he wanted the term to be used. Also about then, D. V. Lindley averred that "There is no one less Bayesian than an empirical Bayesian," a comment that seemed mainly directed at Robbins' approach. Some other statisticians then, and perhaps still today, thought of empirical Bayes as restricted to plugging hyperparameter estimates into Bayes rules. So the term "empirical Bayes" meant different things to different statisticians, and not always good things.

It also had become clear to me back then that dealing with many inferences simultaneously had to be guided by Bayesian reasoning. For example, Bayesian constructions show why interval estimates based on plug-in methods can be much too narrow, especially when the number (N, in the notation of Brad's paper) of random effects being estimated is small or moderate. So I began to use the term empirical Bayes more sparingly to describe my own work. In building on the ideas behind my 1983 paper, and when trying to combine frequency ideas with Bayes in hierarchical models, I sometimes have referred to a "general model for statistics" for the desired frequency/Bayes unification.

The general model includes distributions for data given parameters of interest, and for the hyperparameters that govern the distribution of those parameters, conceptually (but not always) specified for at least two hierarchical levels. From the frequency perspective in this general model, all possible distributions would be considered for the hyperparameters, those being mixtures of atomic (Dirac) distributions. From the subjective Bayesian perspective, just one distribution (a prior at the top level of the hierarchy) would be allowed in a particular inferential problem. (This framework extends to nonparametric models by letting the parameters and/or hyperparameters be infinite dimensional.)

This general model puts frequency and Bayesian models at the endpoints of a continuum, with the middle span open for flexibly specifying restrictions on distributions that could accommodate empirical Bayes and other models. Decision theory extends to this general model so that frequency (resampling) evaluations would be done conditionally for the range of hyperparameters. Such resampling was carried out when evaluating the coverage probabilities of parametric empirical Bayes interval estimates in Morris (1983) and in much other work since then. In a University of Texas dissertation, Joe Hill showed how this general framework extended to ancillarity, information, and other fundamental statistical ideas (Hill, 1986, 1990).

Aside from their different interpretations, the frequency and Bayesian perspectives can be quite complementary. The frequency paradigm is normative, but not necessarily prescriptive. The fundamental theorem of (frequency) decision theory, that is, the complete class theorem, supports the Bayesian connection by recognizing that the admissible procedures nearly coincide with the class of extended Bayes rules. Statistical procedures with good repeated sampling (frequency) properties often can be anticipated by thinking about Bayesian constructions.

A reminder of how Bayesian procedures can have better frequency properties than those derived solely by frequency reasoning is illustrated by a graph with $N = 15$ in Christiansen and Morris (1997, Figure 1). Poissonly distributed summary data like those seen at heart transplant hospitals are fitted there via two-level models. The graph there shows the coverage rates in repeated sampling of nominal 95% intervals when the transplant success rates are simultaneously estimated

at the different hospitals. Six procedures are evaluated. Two follow Bayesian constructions, one that uses the BUGS program and default prior, and the other being an accurate approximation of a hierarchical Bayes procedure based on a hyperparameter prior akin to Stein's superharmonic prior for Normal distributions. These two Bayesianly motivated interval procedures cover or nearly cover 95% of the time in repeated sampling simulations, as intended. The four frequency procedures based on MLE, REML and on two GLM multilevel techniques, have coverages in the range of 60% to 90%, falling well below the claimed coverage rate of 95%. Whether developed from Bayesian or frequency considerations, good frequency procedures must provide coverages in repeated sampling close to their claimed values, but the four non-Bayesian procedures do not meet that standard.

2. FDR, FDR AND EXCHANGEABILITY

Brad illustrates the use of Bayesian modeling and probabilistic reasoning with his six large datasets to produce approaches to hypothesis testing that would be valid if prior information were available. Then he shows how to estimate the needed prior, or mixing, distributions from repeated data.

Probabilistic modeling leads directly to Efron's local fdr, which in turn leads to the Benjamini–Hochberg Fdr procedure. Starting with the simplest "two-groups" model, with density f_0 under the null hypothesis H_0 and f_1 under the alternative hypothesis H_1, the paper moves through increasingly elaborate probability models discovered in the process of modeling and analyzing exchangeable data and repeated problems. Benjamini and Hochberg's celebrated false discovery rate statistic Fdr applies when all the H_0 distributions have a single theoretically determined density function f_0, and when the prior probability p_0 of H_0 is high (at least 0.9). Then f_1, the H_1 density, is available via estimating the marginal density, $f(z) = p_0 * f_0(z) + p_1 * f_1(z)$ and solving for $f_1(z)$. While f_1 is not actually needed in exchangeable cases, it will be for a nonexchangeable extension which I will review later. Thus, a direct estimate of the posterior probability of H_0, given the data, only requires p_0, f_0 and $f(z)$ in this simplest case.

This approach is beguilingly simple, but its validity depends crucially on a restrictive exchangeability assumption that can be missed. The marginal density $f(z)$ will be the same for all the z_i observations only if the same f_1 distribution holds under H_1 for all $z_i, i = 1, \ldots, N$. This may hold for five of the six datasets in the paper, but it does not for the school data, as discussed later.

As formula (2.7) shows, the local fdr is the posterior probability of H_0, that is,

$$\text{fdr}(z) = P(H_0|Z = z) = \frac{p_0 * f_0(z)}{p_0 * f_0(z) + p_1 * f_1(z)}.$$

Starting with fdr(z) before introducing Fdr(z) seems natural, but this particular history has developed oppositely. Efron's local fdr is immediately interpretable in probabilistic or Bayesian terms because choosing between hypotheses H_0 and H_1 means considering $P(H_0|z)$, and also because fdr depends on the likelihood ratio, and on the Neyman–Pearson statistic.

As Brad writes, the Benjamini–Hochberg Fdr statistic (2.3) is the integral of fdr(z). Starting with fdr$(Z) = P(H_0|Z)$ and assuming that one will choose H_1 whenever $Z \le z$ leads to

$$E\big(\text{fdr}(Z)|Z \le z\big) = P(H_0|Z \le z) = \text{Fdr}(z),$$

as shown in the paper, and this is

$$\text{Fdr}(z) = \frac{p_0 * F_0(z)}{p_0 * F_0(z) + p_1 * F_1(z)}.$$

Thus, Fdr$(z) = P(H_0|Z \le z)$ is the fraction of times that H_0 would be falsely rejected. The Benjamini–Hochberg false discovery rate Fdr(z) is discovered probabilistically as the average probability (the preposterior probability in Bayesian terms) of accepting, that is, discovering, H_1 falsely.

The probability model that leads to the fdr and Fdr statistics in repeated applications assumes exchangeability in two ways. First, p_0 should not depend on i, as Efron discusses in Section 2. Second, f_0 and f_1 must be the same for all problems $i = 1, 2, \ldots, N$. From the two-level modeling perspective of the paper, $f_1(z)$ is a mixture of densities for the (approximately) $N * p_1$ values of μ_i that are distributed according to H_1. Denoting the random effects as μ_i for $i = 1, \ldots, N$, exchangeability permits the conditional densities $f(z_i|\mu_i)$ for z_i to depend on i through μ_i only, and not otherwise to depend on i.

Some two-level settings are modeled with "paired" exchangeability among individuals [i.e., the collection of pairs (z_i, μ_i) are exchangeable], and that produces exchangeability for the marginal distributions of z_i. This happens familiarly with N independent individuals (in the paper, "individuals" are the N genes, and the schools, etc.) if the joint distributions of (z_i, μ_i) are i.i.d. Robbins' original introduction of empirical Bayes

for Poisson models rested on paired exchangeability because every individual Poisson distribution was assumed in his paper to have the same exposure. The James–Stein estimator arises as a parametric empirical Bayes estimator, but only when paired exchangeability holds, as when the sample means all have the same variances.

A happy consequence of pairwise exchangeability is that Bayesian procedures often can conveniently be expressed explicitly in terms of the marginal (unconditional) distribution of the data (z_i), and that marginal can be estimated directly from the observed z_i, as Efron has done in several settings. This gives an asymptotically consistent estimate of a Bayes procedure, and the statistician then can avoid directly estimating the mixing distribution $g(\cdot)$ that governs the random effects, μ_i. Relatively simple expressions then may emerge, such as the procedures of Robbins, of Stein, and of Benjamini–Hochberg. As Efron notes, the independence assumption is not crucial, but exchangeability is. The Fdr and fdr statistics in the exchangeable setting of Efron's Section 2 should work well with pairwise exchangeability when N is large, but exchangeability can be restrictive and may depend heavily on prior knowledge. Seemingly, exchangeability is widely considered to hold for microarray, proteomics, BRCA and spectroscopy data. It cannot be valid for the school data because school enrollments, that is, sample sizes n_i vary. Nearly all theory presented in this paper is based on such exchangeability, barring the discussion of nonexchangeable choices for p_0 in Section 2. Is "empirical Bayes" in this paper meant to be limited to exchangeable (or pairwise exchangeable) settings?

3. MULTIPLE HYPOTHESIS TESTING—LOOKING FOR LARGE RANDOM EFFECTS

Here is an extension of Efron's approach that may be especially useful for identifying large random effects μ_i. First consider and fix any single value of i, $1 \le i \le N$, with $z = z_i$ having been observed, and assume that the "theoretical null" $N(0, 1)$ distribution holds for z_i under H_0, that is, when the random effect $\mu = \mu_i = 0$. Assume p_0, f_0 and f_1 all are known for this value i, as in Section 2, and that $g(\cdot)$ is known. Then $f_1(z)$, the marginal distribution of z under H_1, is determined by integrating the conditional distribution of z given μ, for example, $z \sim N(\mu, 1)$ having density $\phi(z - \mu)$, over the distribution $g(\mu)$ that governs the H_1 distribution of μ. (Exchangeability does not matter when all these distributions are known.) Then when H_1 holds, the density of μ given z is

$$h(\mu|z) = \phi(z - \mu) * g(\mu)/f_1(z).$$

With $\text{fdr}(z) = P(\mu = 0|z)$, and writing $\delta(\mu)$ as the Dirac delta function ($\mu = 0$ with probability 1 when H_0 is true), the density of μ given z is expressible as a mixture of Efron's $\text{fdr}(z)$ according to

$$p(\mu|z) = \text{fdr}(z) * \delta(\mu) + \big(1 - \text{fdr}(z)\big) * h(\mu|z).$$

If all these distributions and values were known, one could "test" $H_0 : \mu = 0$ (or $\mu \le 0$?) versus $H_1 : \mu > 0$ by using $\text{fdr}(z)$ as the probability of H_0. However, one well might prefer only to identify genes "far from H_0," that is, only select values of $\mu > k$ that exceed a scientifically substantial magnitude $k > 0$, and with a substantial probability. One then would use $p(\mu|z)$ in the formula above to calculate $P(\mu \ge k|z)$.

Numerical illustrations are easy to do, and here is one based on the assumptions in Section 5 of the paper, with $N = 3000$, $p_0 = 0.9$, and Normal distributions with $z_i \sim N(\mu_i, 1)$ and $g(\mu_i)$ being the $N(2.5, 0.5)$ density. Then values of $z \ge 3.5$ occur in 2.1% of the genes, so $z \ge 3.5$ identifies about 63 of the 3000 genes. If we were to choose $k = 2.8$, then $P(\mu > 2.8|z) = 0.506$ at the threshold value $z = 3.5$, and the conditional probability that $\mu > 2.8$ rises as z increases. Researchers who wish to identify about 63 genes (2.1%) would calculate $P(\mu_i > 2.8|z_i)$ for every one of the 63 selected genes, all those that have at least a 50% chance of $\mu > k = 2.8$, and (by averaging) that overall about 60% of the 63 selected cases have $\mu_i > 2.8$. The 60% statement is analogous to Benjamini–Hochberg's calculation, calculated here by averaging the 63 selected posterior probabilities. If a smaller value $k = 2.0$ were chosen, then selected genes at that threshold, still with $z \ge 3.5$, would have at least a 90% chance (90% if $z = 3.5$ exactly) that $\mu > 2.0$, and one would know that about 95%, or 60 of the 63 selected cases, would have $\mu > 2.0$. Of course, if $k = 0$, as in the paper, then $\text{fdr}(z)$ and $F(z)$ would indicate that about 98% (61 or 62) of the 63 selected cases with $z \ge 3.5$ would have $\mu > 0$.

The preceding assumes a one-tailed test, as does Fdr, and so we have used $k > 0$ (if large positive values of μ are wanted), but two-tailed probabilities also are easy to evaluate. A table of the $N = 3000$ genes could list genes, sorted by their values of $P(\mu > k|z)$, using $p(\mu|z)$. With exchangeability, the ordering is that of z_i. Researchers could review these values of $P(\mu > k|z)$, keeping as many genes as desired, and stop when this probability becomes too low, or when

enough candidates have been accepted. There is nothing special about keeping 2.1% and changing the cutoff for z would alter that percentage. Experience gained with different values of k after a variety of analyses with various data sets eventually might help identify the scientifically most useful values.

Of course $g(\mu)$ and the other constants are not generally known. That is the point of Efron's paper, but g can be estimated by a variety of methods, frequentist, Bayesian and empirical Bayesian, and perhaps quite accurately with large N. The paper shows some nifty ways to estimate f_1 in exchangeable settings. Then one could use the estimated f_1 to estimate $g(\mu)$, perhaps by deconvolution methods. While estimating these mixing distributions $g(\cdot)$ becomes more difficult in nonexchangeable cases when the z_i have different conditional distributions given μ_i, the literature provides a variety of ways to do that, most easily in parametric settings.

The proposal just described would test interval null hypotheses instead of single points by calculating $P(H_1)$ given the data, also by using the data to learn about various constants and distributions, for example, about $p_0, g(\cdot)$, etc. Doing this in conjunction with choosing a $k > 0$ has been recommended in medical profiling by Burgess, Christiansen, Michalak and Morris (2000) for profiling hospital performances. Standard practice for medical profiling most commonly is based on testing different hypotheses like $H_0 : \mu_i = 0$ independently, using standard methods like those widely taught in beginning statistics courses. That forfeits the possibility of developing more information via multilevel modeling. Once multilevel models have been fitted, it is natural to consider alternative hypotheses like $H_1 : \mu > k$ where $k > 0$ is chosen to set standards (k) for unacceptable (or laudatory) departures from average outcomes of medical procedures. The analogous proposal is made here, which can be extended to accommodate a spike at 0 with $p_0 > 0$ within $H_0 = (-k, k)$ if required. That extension is not needed with medical profiling data, where it is unlikely that any sizeable fraction p_0 of hospitals would have precisely the same underlying rates of surgical outcomes, but the paper's applications make it clear that positive probability for a null point within H_0 is appropriate in a variety of problems.

In exchangeable cases, ranking according to p-values will not depend on the choice of k. With medical outcome data for hospitals, the number of treated patients always will vary substantially, producing nonexchangeability. Then shrinkages toward a common mean will be greater for small hospitals than for large ones, and the resulting rankings will depend not only on z_i, but also on n_i and on k.

4. NONEXCHANGEABILITY, THE SCHOOL DATA AND THE ONE-GROUP MODEL

The school data of Figure 1(b) are not exchangeable because the sample sizes n_i (actually there are two different sample sizes for each school, one for each demographic group) surely vary across the $N = 3748$ schools. Equal sample sizes might lead to exchangeability, but that rarely happens except with designed experiments, as the microarray experiments must be. Together (e.g., in Efron and Morris, 1975), Brad and I once used toxoplasmosis summaries for $N = 36$ regions to illustrate generalizations of Stein's estimator that were needed to account for different sample sizes in different regions. Those toxoplasmosis data, the hospital profiling data, and the school data in this paper all might be similarly modeled. The school data calculations suggest shrinkages should vary, but average about 40%. A sharp null with p_0 much in excess of 0 seems implausible for toxoplasmosis, for hospital data, and for the school data, and so Efron introduces the case $p_0 = 0$ as his "one group model." One would then expect that $\mathrm{Var}(z_i)$ is proportional to n_i. That would cause longer-tailed distributions for the $\{z_i\}$ values than Normality allows, and schools with more students would tend to be the outliers. Figure 1(b) reveals evidence of such long tails, corresponding to nonexchangeability.

5. INTERVAL ESTIMATION

Efron's Section 7, about interval estimation, shows in a simulation with exchangeable data that the FCR intervals are too wide. That happens because the FCR does not adjust its slope to be less than 1.0 when a gentler slope closer to 0.5 would track regression toward the mean (RTTM) of the 1000 random effects better. Interval estimates recentered according to this slope improvement can be shorter and still cover at the same rate as FCR does. Morris (1983) provides a basis for evaluating interval coverages via repeated sampling. Figure 1 in that 1983 paper (data for $N = 18$ baseball players from some early Efron–Morris papers) illustrates how intervals centered on shrunken estimates are much more accurate. The graph there makes the same point that Efron does in Figure 8. However, Brad's Section 7 conclusion avers that Bayesian intervals cannot be trusted. That does not square with my experience because I have found Bayesian reasoning to be

essential to understanding how to construct interval estimates that have good frequency properties.

With 1000 observations it makes sense to estimate the distribution $g(\mu)$ without assuming Normality, and instead to use exchangeability as a basis for estimating the marginal distribution of the $\{z_i\}$. The same can be done with Bayesian methods, even with a nonparametric specification for g, although less easily. With unequal sample sizes, or when N is not large, a Bayesian approach may be more successful, as with the heart transplant data of Christiansen and Morris (1997). A key to knowing that Bayesianly constructed confidence intervals will meet frequency resampling criteria requires identifying and using frequency-friendly noninformative distributions for the hyperparameters. This has been done in a variety of specific parametric settings, including for some common generalized linear models. Bayesian reasoning also shows us how to account for added variability in settings where the hyperparameters and shrinkage constants have been estimated. Such intervals must bow outward in Efron's Figure 8 when moving away from the center, and this is seen more dramatically when $N = 18$ in Figure 1 of Morris (1983). Efron's Figure 8 shows no bowing, but that would be too small to see with such large sample sizes. More discussion is needed as to whether Bayesian reasoning really has failed in the Section 7 setting, and about what an empirical Bayes approach really can offer, beyond suggesting Bayesian methods designed to withstand frequency verifications.

6. MODELING AND RTTM

Two-level modeling can reveal by how much random effects will regress (shrink) toward the mean (RTTM). The modeling task is to estimate the mean to shrink toward, and determine how much shrinkage. A term I always liked that Brad used when we wrote together is "ensemble information." RTTM means individual estimates will regress toward the ensemble estimates.

The paper focuses on rectangular X as an N-by-n data matrix. When X is rectangular, it is especially valuable to analyze the distribution of the rows and columns of X, calculating correlations as Brad does among the rows (genes) and/or among the columns (arrays) to improve estimates of f_0, f_1 and p_0, and thereby to keep modeling assumptions at a minimum.

Of course X need not be rectangular, nor should it automatically be so considered, because different rows sometimes may contain different amounts of data. The school data would follow such a nonrectangular shape if each row were to include separate entries for each student (as the BRCA and the HIV data do, but always with the same sample sizes). In this case, the school data have been forced into a rectangular Procrustean bed by using summarized data z_i, and that has obscured their nonexchangeability.

Sometimes it pays to take advantage of situations when N is large, but without appealing to asymptotics. In the context of the paper, that might be done by increasing the number of parameters and fitting richer models as N increases. This is parametric model-building, but it is an alternative to nonparametric modeling. The paper does some of this to investigate correlations, but the same could be done to assess whether exchangeable models are adequate.

A model for microarray data considered by Hongkai Ji and Wing Wong (Ji and Wong, 2005), concerns dealing with the (nuisance) standard deviations σ_i that are estimated in the denominator of each t-statistic, like those considered in Sections 4 and 5 in Efron's paper. The sample standard deviations s_i (each based on just a few degrees of freedom, as with the BRCA and HIV data) easily can produce randomly small sample standard deviations to estimate σ_i, and hence produce large t-values that falsely indicate which genes are expressing themselves importantly. A way out of this is to consider the N problems to be exchangeable with respect to the random effects μ_i and also the σ_i. That justifies shrinkage methods (based on chi-squared distributions). Ji shows that shrinking the sample standard deviations s_i toward their common mean, and using these empirical Bayes shrunken estimates in place of s_i in the t-statistics, greatly improves the rate of false gene discoveries.

7. CONCLUSION

Brad Efron's paper introduces many ideas for analyzing massive datasets. It encourages a frequency-Bayes unification and empirical Bayes modeling. The paper identifies modeling and inference opportunities that arise with massive datasets in exchangeable settings. Much remains to do to understand the exchangeable case for parametric and nonparametric models alike, and there is much to do to recognize when nonexchangeable models are required, and how to fit them.

REFERENCES

BURGESS, J. F., CHRISTIANSEN, C. L., MICHALAK, S. E. and MORRIS, C. N. (2000). Medical profiling: Improving standards and risk adjustments using hierarchical models. *J. Health Economics* **19** 291–309.

CHRISTIANSEN, C. L. and MORRIS, C. N. (1997). Hierarchical Poisson regression modeling. *J. Amer. Statist. Assoc.* **92** 618–632. MR1467853

EFRON, B. and MORRIS, C. N. (1975). Data analysis using Steinas estimator and its generalizations. *J. Amer. Statist. Assoc.* **70** 311–319. MR0391403

HILL, J. R. (1986). Empirical Bayes statistics: A comprehensive theory for data analysis. Ph.D. dissertation, Univ. Texas.

HILL, J. R. (1990). A general framework for model based statistics. *Biometrika* **77** 115–126. MR1049413

JI, H. and WONG, W. A. (2005). TileMap: Create chromosomal map of tiling array hybridizations. *Bioinformatics* **21** 3629–3636.

MORRIS, C. N. (1983). Parametric empirical Bayes inference: Theory and applications (with discussion). *J. Amer. Statis. Assoc.* **78** 47–65. MR0696849

Statistical Science
2008, Vol. 23, No. 1, 41–44
DOI: 10.1214/07-STS236A
Main article DOI: 10.1214/07-STS236

Comment: Microarrays, Empirical Bayes and the Two-Groups Model

Kenneth Rice and David Spiegelhalter

Through his various examples, Professor Efron makes a convincing case that cutting-edge science requires methods for detecting multiple "non-nulls." These methods must be straightforward to implement, but perhaps more importantly statisticians need to be able to justify them unambiguously. Efron's Empirical Bayes approach is certainly computationally efficient, but we feel the rationale for making each of his steps is unattractively ad hoc. This concern is practical, not philosophical; Efron's criterion for choice of tuning parameters seems to be that they look "believable." In less expert hands, this approach seems to introduce a lot of leeway for practitioners to simply "tune" away until they get the results they want.

In an attempt to address this problem, we will describe an approach developed in a fully model-based framework. As with locfdr, the calculations are fast, but our whole analysis derives from clear up-front statements about what the analysis is trying to achieve, and the modeling assumptions made. The results look reassuringly similar to Professor Efron's. We hope this will be helpful for understanding the current paper, and in making a contribution to this general field.

We begin by following Efron in placing the local false discovery rate, $\text{fdr}(z)$, as the primary focus of the analysis, and exploit the fact that it can offer a neat parameterization of the two-part model. If the marginal, "mixture" density for the z-values is

$$f(z) = p_0 f_0(z) + (1 - p_0) f_1(z)$$

and $\text{fdr}(z) = p_0 f_0(z)/f(z)$, then

$$f_1(z) = \frac{p_0}{1 - p_0} \frac{1 - \text{fdr}(z)}{\text{fdr}(z)} f_0(z).$$

Kenneth Rice is Assistant Professor, Department of Biostatistics, University of Washington, Seattle, Washington 98195-7232, USA (e-mail: kenrice@u.washington.edu). David Spiegelhalter is Senior Scientist, MRC Biostatistics Unit, Institute of Public Health, Cambridge CB2 OSR, United Kingdom (e-mail: David.Spiegelhalter@mrc-bsu.cam.ac.uk).

We observe that, because f_1 is a density, we only need to know f_0 and fdr in order to find its normalized form, and in turn this tells us the value of p_0. Thus, for a given f_0, specifying fdr sets up everything else we require for model-based analysis.

Naturally, the analysis we report will depend on the functional form assumed for fdr, and Efron implicitly assumes a rather flexible form of fdr, through a seventh-order polynomial-smoothed density estimate. However, this approach does not rule out an $\widehat{\text{fdr}}$ with multiple peaks. Thinking of the schools example, we would not want to be the statistician explaining how two "bad" schools may have $z_1 < z_2 < 0$, but yet $\widehat{\text{fdr}}(z_1) > 0.2$ while $\widehat{\text{fdr}}(z_2) < 0.2$. Put more simply, Efron's method can report that School 1 has worse performance, but only School 2 is called an outlier. We find it more straightforward to a priori justify our choice of fdr by careful consideration of its role in the reported inference.

In our experience, the search for non-null "discoveries" is based around two ideas; first, we will not discover anything near the center of f_0 (effectively Efron's "zero assumption," also termed "purity" by Genovese and Wasserman, 2004). A second sensible assumption is that the evidence for z being "null" will decrease monotonically as we move out from the center. One way to satisfy this is with a logistic-linear form for fdr, giving a two-component normal mixture for f_1, but we get closer to the spirit of Efron's analysis by assuming that fdr is unity inside a central region, and then follows a half-normal decline, that is,

$$\text{fdr}^H(z) = \begin{cases} e^{-(z+k_a)^2/2}, & z < -k_a, \\ 1, & -k_a \le z \le k_b, \\ e^{-(z-k_b)^2/2}, & z > k_b. \end{cases}$$

Following the observation above, taking the null component f_0 to be standard Normal, now defines the following marginal distribution $f^H(z)$:

$$f^H(z) = p_0(2\pi)^{-1/2} \cdot \begin{cases} e^{-|z|k_a + k_a^2/2}, & z < -k_a, \\ e^{-z^2/2}, & -k_a \le z \le k_b, \\ e^{-|z|k_b + k_b^2/2}, & z > k_b, \end{cases}$$

where the constant of proportionality is p_0, the proportion of nulls, which is an easily determined function of k_a and k_b.

$f^H(z)$ is seen to have a $N(0, 1)$ "core" and exponential tails. By substituting $(z-\mu_0)/\sigma_0$ for z in $f_0(z)$ and $\text{fdr}^H(z)$, it is easily generalized to a full location-scale family, where the "core" (or null distribution) is now $N(\mu_0, \sigma_0^2)$. We term this a "Huber" distribution, denoted $H(\mu_0, \sigma_0, k_a, k_b)$, following the observation in Huber (1964) that his optimal robust location estimation procedure based on a piecewise-linear bounded influence function was precisely equivalent to maximum likelihood estimation applied to such a distribution, but with $k_a = k_b = k$ specified and σ_0 assumed known.

Assuming this distribution and adopting a full likelihood approach, maximum likelihood estimates $\hat{\mu}_0, \hat{\sigma}_0$ are the solutions of estimating equations that take, up to a very good approximation, the same form as Huber's famous "Type 2" estimator. We do not need to fix k_a and k_b; they can be estimated from the data in the same way.

We have implemented maximum likelihood-based regression for this error distribution within our own R package (*huber.lm*), and also as a fully Bayesian MCMC approach via a new distribution, *dhuber*, within WinBUGS.

Figure 1 and Table 1 show the results of fitting this distributional family to four of Efron's examples using *huber.lm*.

In line with Efron, we assume that f_0 follows a $N(\mu_0, \sigma_0^2)$ distribution, and provide point estimates for μ_0, σ_0, p_0 as well as k_a, k_b. We also show the fitted marginal distributions $f^H(z)$, QQ-plots of the z-values against $f^H(z)$ and a "naive" Normal, the fitted local false discovery rate $\text{fdr}^H(z)$, and an appropriately scaled representation of the "alternative" distribution f_1. Figure 1 shows a good fit of the Huber distribution to these examples. The fitted fdr^H curves are also plotted, and these show a close concordance with Efron's locfdr results. For the BRCA data, we have not plotted fdr^H, as use of the Huber distribution here gives estimates for both k_a and k_b tending to ∞, and hence gives a point estimate of $\widehat{\text{fdr}} = 1$ for all data points. The practical message is clear; we find that the BRCA data, on its own, provides no strong evidence of any signals beyond the fitted $N(\mu, \sigma^2)$ null, in line with Efron's results. The QQ-plot for the BRCA data provides further informal confirmation. Other authors have declared some evidence for signals in this dataset, a recent example being Jin and Cai (2007). However, this is in contrast to a Bayesian analysis with a uniform prior for k_a and k_b, which leads to a posterior for both k_a and k_b that rules out values less than 2 ($p_0 > 0.8\%$) and which provides an essentially uniform distribution for $k_a, k_b > 3$ ($p_0 < 0.02\%$).

Table 1 provides parameter estimates for the asymmetric Huber distribution: likelihood ratio tests for common k are $p = 0.68$ (Prostate); $p = 0.14$ (Education); $p = 0.007$ (HIV). We find a close concordance between our results and those in Efron's paper. The estimated proportions of nonnull observations are 1.7% (Prostate), 7.3% (Education) and 6.2% (HIV). As p_0 is a slightly messy function of k_a and k_b,

$$p_0 = \sqrt{2\pi}\big[e^{-k_a^2/2}/k_a + e^{-k_b^2/2}/k_b + \sqrt{2\pi}\big(\Phi(k_a) + \Phi(k_b) - 1\big)\big]^{-1},$$

we have found it easiest to obtain intervals by using an MCMC approach. However, using the delta method or a parametric bootstrap on the distribution of the MLEs offers, in spirit, the same inference.

In contrast to Efron's desire to "minimize the amount of statistical modeling required of the statistician," we would encourage statistical modeling *where the modeling assumptions are clear and comprehensible*; for example, we find a simply defined parametric model preferable to Efron's seven-parameter polynomial-smoothed density estimate. Our explicit acknowledgment of these assumptions also motivates consideration (below) of how they may be usefully strengthened, and also whether they may be relaxed.

Using a simple but flexible fully parametric family such as the Huber distributions confers many advantages. If we are willing to condition on the adequacy of the assumed model for $f^H(z)$, then the full resources of likelihood modeling become available, providing interval estimates, hypothesis tests and so on. In a hierarchical setting, the Huber distribution can also be considered at the random-effects level. Computationally this is handled with ease within a full Bayesian MCMC environment, where using $H(\mu, \sigma, k)$ or $H(\mu, \sigma, k_a, k_b)$ within a hierarchical model presents no additional difficulties over its use as a sampling distribution. Becoming "more" Bayesian still, we note the possibilities for use of informative priors regarding the thresholds k_a and k_b, and hence implicitly p_0. In our opinion, analyses which acknowledge these a priori assumptions seem particularly attractive for examples smaller than Efron's, where a reliable density estimate seems out of reach. Finally, a Bayesian modeling framework allows the inclusion of a model for such data within an integrated evidence

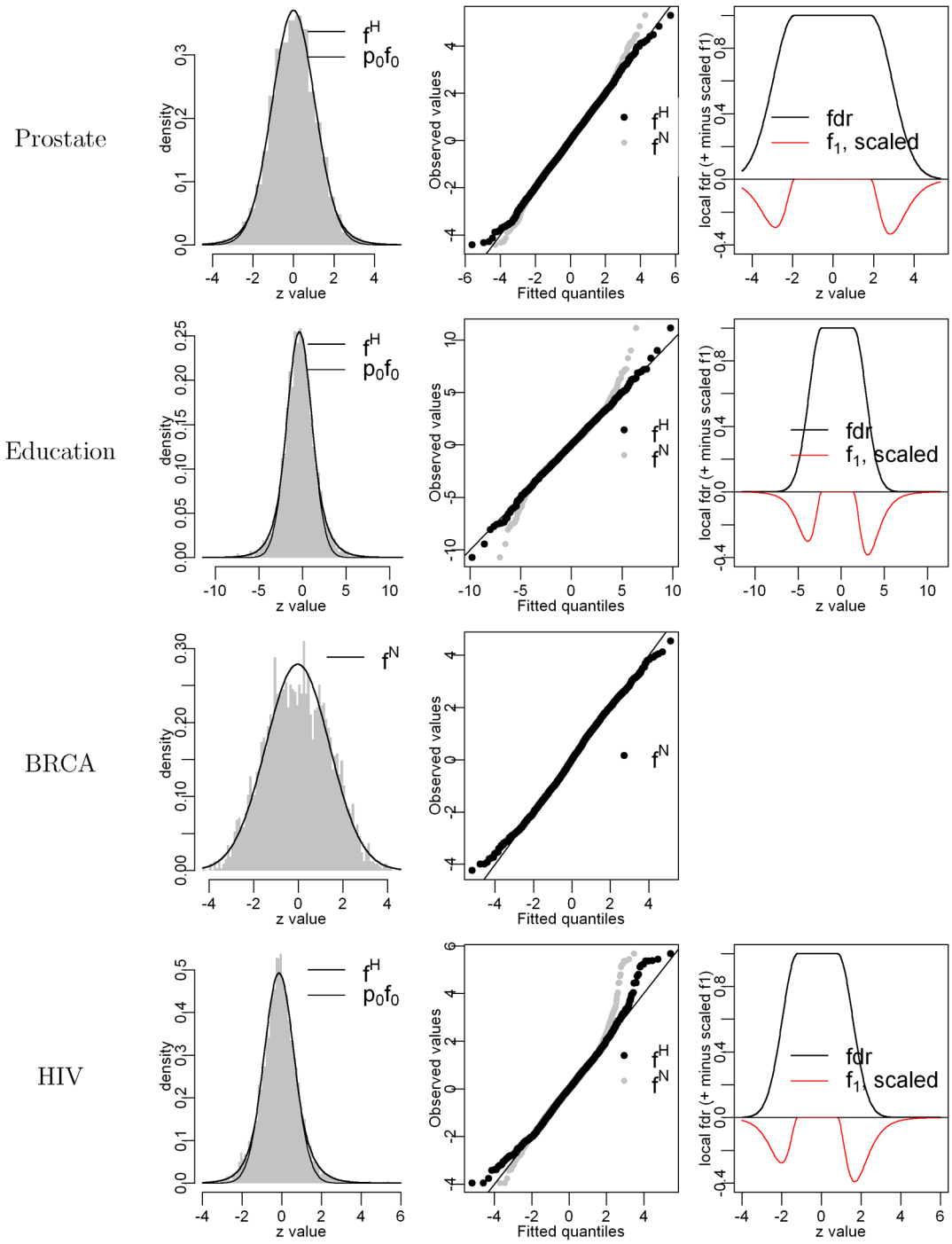

FIG. 1. *Summary plots fitting the Huber distribution to four examples. For each dataset, we plot histograms of the z-values and fitted marginal distribution, QQ-plots of the data against fitted Huber distribution* (f^H) *and a naive pure Normal* (f^N), *and finally a plot of the fitted* fdr *and the alternative distribution* f_1 (*inverted*). *For BRCA, the fitted* fdr *is always* 1, *giving no strong evidence of signals in this dataset.*

TABLE 1
Maximum likelihood estimates (95% *intervals*) *for parameters of the asymmetric Huber distribution for four of Efron's examples*; *the intervals for* p_0 *are obtained from an MCMC simulation*

	Prostate		Education		BRCA		HIV	
μ_0	−0.001	(−0.031, 0.030)	−0.361	(−0.427, −0.295)	−0.026	(−0.075, 0.023)	−0.138	(−0.161, −0.115)
σ_0	1.059	(1.030, 1.089)	1.452	(1.363, 1.546)	1.431	(1.396, 1.466)	0.760	(0.730, 0.791)
k_a	1.80	(1.61, 2.01)	1.31	(1.17, 1.48)	—	—	1.40	(1.28, 1.53)
k_b	1.75	(1.59, 1.93)	1.21	(1.08, 1.37)	—	—	1.26	(1.17, 1.36)
p_0	0.983	(0.975, 0.990)	0.927	(0.899, 0.950)	—	—	0.938	(0.921, 0.954)

synthesis, which can be guided by a combination of substantive knowledge and data analysis.

Taking a less Bayesian or full-likelihood approach, and not wishing to condition on the "truth" of the model assumptions, one could proceed directly to Huber-style estimating equations for μ_0, σ_0 and k (or k_a and k_b), justified either through their connection to the model we have described, or by arguing that this influence function directly reflects the population parameter we want to estimate; if we are trying to minimize model-dependence, the second approach is more satisfactory, and is quite standard in GEE. Sandwich and/or bootstrap variance estimates could be used to reflect uncertainty about these point estimates, without further parametric assumptions about the mixture distribution f. In samples of thousands of z's (but not with a few hundred), this provides appealingly robust estimates of location and scale.

However, going beyond μ_0 and σ_0, it is not clear to us that the GEE paradigm allows "model-robust" measures of fdr. Must one compare the marginal f to an f_0 which is assumed to have a specifically Gaussian form, or that of some other parametric family? Might some advanced form of cross-validation offer a model-free approach? And could this be done without an excessive computational burden? Any insights from Professor Efron in this matter would be very welcome.

In conclusion, we feel that flexible likelihood or Bayesian modeling techniques, combined with basic insights from the literature on outlier-robustness, will contain much of value in the era of microarrays and other data-sources requiring large numbers of hypothesis tests. We thank Professor Efron for his stimulating paper, and also for his generosity in making available the four featured datasets.

REFERENCES

GENOVESE, C. and WASSERMAN, L. (2004). A stochastic process approach to False Discovery Rates. *Ann. Statist.* **32** 1035–1061. MR2065197

HUBER, P. J. (1964). Robust estimation of a location parameter. *Ann. Math. Statist.* **35** 73–101. MR0161415

JIN, J. and CAI, T. (2007). Estimating the null and the proportion of nonnull effects in large-scale multiple comparisons. *J. Amer. Statist. Assoc.* **102** 495–506. MR2325113

Statistical Science
2008, Vol. 23, No. 1, 45–47
DOI: 10.1214/08-STS236REJ
Main article DOI: 10.1214/08-STS236

Rejoinder: Microarrays, Empirical Bayes and the Two-Groups Model

Bradley Efron

The Fisher–Neyman–Pearson theory of hypothesis testing was a triumph of mathematical elegance and practical utility. It was never designed, though, to handle 10,000 tests at once, and one can see contemporary statisticians struggling to develop theories appropriate to our new scientific environment. This paper is part of that effort: starting from just the two-groups model (2.1), it aims to show Bayesian and frequentist ideas merging into a practical framework for large-scale simultaneous testing.

False discovery rates, Benjamini and Hochberg's influential contribution to modern statistical theory, is the main methodology featured in the paper, but I really was not trying to sell any specific technology as the final word. In fact, the discussants offer an attractive menu of alternatives. It is still early in the large-scale hypothesis testing story, and I expect, and hope for, major developments in both theory and practice.

The central issue, as Carl Morris makes clear, is the combination of information from a collection of more or less similar sources, for example from the expression levels of different genes in a microarray study. Crucial questions revolve around the comparability and relevance of the various sources, as well as the proper choice of a null distribution. Technical issues such as the exact control of Type I errors are important as well, but, in my opinion, have played too big a role in the microarray literature. The discussions today are an appealing mixture of technical facility and big-picture thinking. They are substantial essays in their own right, and I will be able to respond here to only a few of the issues raised.

I once wrote, about the jackknife, that *good* simple ideas are our most precious intellectual commodity. False discovery rates fall into that elite category. The two-groups model is used here to unearth the Bayesian roots of Benjamini and Hochberg's originally frequentist construction. In a Bayesian framework it is natural to focus on local false discovery rates, fdr(z), rather than the original tail area version Fdr(z). My apologies to Professor Benjamini for seeming to suggest that fdr is more immune than Fdr to correlations between the z-values. All false discovery rates are basically ratios of expectations, and as such remain relatively unbiased in the face of correlation. It is only the proof of the exact Fdr control property that involves some form of independence.

In the same spirit, I have to disagree that Fdr produces more reproducible results than fdr. Both methods operate at the mercy of an experiment's power, and low-power situations, such as the prostate cancer study, are certain to produce highly variable lists of "significant" cases. (At this point, let me repeat my plea for a better term than "significant" for the cases found to be nonnull, a dubious nomenclature even in classical settings, and definitely misleading for large-scale testing.)

As suggested by Figure 2, there is no great conceptual difference between fdr and Fdr, nor have I found much difference in applications. Table 1 says something about their comparative estimation accuracy. As Professor Cai suggests, the statistician can combine the two, using Fdr to select a reportable list of nonnull candidates, and fdr to differentiate the level of certainty within the list. Here the two roles reflect Benjamini's distinction between decision theory and inference, that is, between making a firm choice of nonnull cases and providing an estimate of just how nonnull they are.

As an enthusiastic collector of reasons to distrust the theoretical null distribution, I am happy to add *preselection of cases* to the list. Professor Benjamini correctly points out the dangers of this practice—among other things, it deprives the statistician of crucial evidence about the null distribution. If questioning the theoretical null seems heretical, it is worth remembering similar questions arising in classical ANOVA applications, for instance whether to use σ^2 (error) or σ^2 (interaction) in assessing the main effects of a two-way table. I share Benjamini's preference for finding the "right" theoretical null, but that is the counsel of perfection, often unattainable in examples like the education data.

Bradley Efron is Professor, Department of Statistics, Stanford University, Stanford, California 94305, USA (e-mail: brad@stat.stanford.edu).

Questions of exchangeability play a key role in large-scale hypothesis testing, as emphasized in Professor Morris's nice essay. The answer to "Which problems should be tested together?" is not always "All the ones the investigator put on my desk." A paper written after this article, "Simultaneous inference: When should hypothesis testing problems be combined?" (Efron, 2008) attacks this problem without conquering it. As Morris points out, covariates like school size in the education example may undercut exchangeability—the nonnull z-values for larger schools might lie farther away from 0. My paper suggests how to incorporate covariates into an efficient fdr analysis.

In this paper, only the paragraph following that of (3.10) has anything to say about exchangeability. (Notice that the local fdr puts less strain on exchangeability than tail-area Fdr since only the cases near some particular value z are considered together.) For the education example we might be willing to accept exchangeability for the null z_i's, from simple binomial calculations, though not for the nonnull cases. The interpretation of the equivalent of "2.68/17" in the paragraph following (3.10) could thus be modified in a Bayesian way to assign greater nonnull probability to the larger schools.

Morris' Section 3 is especially pertinent. His formula for $p(\mu|z)$ is related to my discussion of the Benjamini–Yekutieli False Coverage Rate method in Section 7, particularly (7.2)–(7.4). Originally I had hoped to develop an empirical Bayes method for estimating such models, but the effort foundered on practical difficulties involving the perils of deconvolution.

Section 6 on the "one-group model" is the ugly duckling of the current paper, but it bears on some important points raised in the discussion. Figure 7 concerns a fuzzy version of simultaneous hypothesis testing, where, as in Morris' hospital example, the usual single-point null hypotheses seem unequal to the task. The development from (6.6) onwards, particularly (6.12), bears on the possibility of nonnormal null distributions, and is about as far as I can go in answering Professors Rice and Spiegelhalter's penultimate question.

With $g(\mu)$ a normal distribution, model (6.1) returns us to the realm of Stein estimation, the scene of my happy collaborations with Carl Morris. I continue to be surprised at how much harder bumpy, nonnormal models like (7.1) are to deal with. James–Stein estimation works fine with, say, $N = 10$ component problems, while the Robbins' form of empirical Bayes appropriate to (7.1) seems to require hundreds or thousands. The information calculations in Efron (2008) reinforce this gloomy assessment. Maybe I am trying to be overly nonparametric in constructing the empirical Bayes Fdr estimates, but it is hard to imagine a generally satisfactory parametric formulation for (6.1). Or perhaps it is just that hypothesis testing is more demanding than estimation.

Rice and Spiegelhalter propose an attractive algorithm: rather than modeling the marginal density $f(z)$ as in (3.6), they suggest directly modeling $\mathrm{fdr}(z)$. The resulting Huber form for $f(z)$ has a pleasant appearance, and I was relieved to see their results agreeing with mine.

The Rice–Spiegelhalter model involves only two free parameters, k_a and k_b, as opposed to seven in (3.6). I doubt that two will be enough to cover a general range of applications, but would be happy to be proved wrong. For example, it might sometimes be necessary to have different exponential rates of decay in the two tails, rather than forcing them to be the same. [Perhaps I am just trying to lob the "ad hoc" accusation back into Rice and Spiegelhalter's court. Equations (3.4)–(3.6) describe a standard Poisson regression model; users of *locfdr* can select the degree of the regression, seven being only the default.] In any case, the direct modeling of $\mathrm{fdr}(z)$ is a promising new route of attack.

"Efficiency" in Professor Cai's essay is what I called "power" in Section 3, a somewhat neglected aspect of multiple testing that now seems to be attracting attention. My diagnostic $E\widehat{\mathrm{fdr}}^{(1)}$, (3.9), is trying to estimate the power parameter

$$1 - \int \mathrm{fnr}(z) \cdot f_1(z)\, dz,$$

where $\mathrm{fnr}(z)$ is the "local false nondiscovery rate" $1 - \mathrm{fdr}(z)$, to use Cai's terminology. See Efron (2007).

Usually $\widehat{\mathrm{fdr}}(z)$ declines monotonically as we move away from $z = 0$ in either direction, so that in each tail $\widehat{\mathrm{fdr}}(z_i)$ orders evidence against the null in the same way as the p-value, p_i. The ordering can be different, however, if we try to compare evidence across the two tails. Cai's results, with Sun, show that it is better to define the decision boundary in terms of fdr-values than p-values, for example by $\widehat{\mathrm{fdr}}(z_i) \le 0.2$ rather than using a p-value cutoff. This nicely reinforces the utility of the Bayesian quantity $\mathrm{fdr}(z)$ (2.7) for frequentist decision-theoretic calculations.

Jin and Cai have a quite different method for empirical null estimation, based on Fourier analysis. This moves in the opposite direction from Rice and Spiegelhalter, more nonparametric rather than less, and again seems to give good estimates.

Large-scale statistical inference blurs the line between Bayesians and frequentists: Bayesian information accumulates, and cannot be ignored, but the accumulation itself favors the use of frequentist tactics. The definition of "empirical Bayes," if there is one, lies somewhere in the realm of Bayesian–frequentist cooperation. Morris points out that this broad-sense definition of empirical Bayes was too wide for Robbins, and maybe for him too, but it is probably enough for the methodological goals of this paper.

My thanks go to the discussants, and also to the editor Ed George for organizing a session on this lively topic.

REFERENCES

EFRON, B. (2007). Size, power, and false discovery rates. *Ann. Statist.* **35** 1351–1377. MR2351089

EFRON, B. (2008). Simultaneous inference: When should hypothesis testing problems be combined? *Ann. Appl. Statist.* **2** 197–223.