Doing Thousands of Hypothesis Tests at the Same Time

Bradley Efron

Abstract

The classical theory of hypothesis testing was fashioned for a scientific world of single inferences, small data sets, and slow computation. Exciting advances in scientific technology – microarrays, imaging devices, spectroscopy – have changed the equation. This article concerns the simultaneous testing of thousands of related situations. It reviews an empirical Bayes approach, encompassing both size and power calculations, based on Benjamini and Hochberg's False Discovery Rate criterion. The discussion precedes mainly by example, with technical details deferred to the references.

Keywords Empirical Bayes, False Discovery Rate, Two-groups model, local fdr, locfdr, microarray, power.

1. Introduction These are exciting times for statisticians. New "high throughput" scientific devices – microarrays, satellite imagers, proteomic chips, fMRI scanners – permit thousands of different cases to be examined in a single experimental run. What arrives at the statistician's desk is often a huge matrix comprising data from thousands of simultaneous hypothesis tests.

Figure 1 concerns a microarray example. Expression levels for $N = 6033$ genes were obtained for $n = 102$ men, $n_1 = 50$ normal subjects and $n_2 = 52$ prostate cancer patients (Singh et al., 2002). Without going into biological details, the principal goal of the study was to pinpoint a small number of "non-null" genes, that is, genes whose levels differ between the prostate and normal groups.

In this case the data matrix X is 6033 by 102, row i being the expression levels for gene i, and column j for microarray j, with the first 50 columns representing the normal subjects Row i provides the usual 2-sample t-statistic "t_i", 100 degrees of freedom, comparing prostate cancer and normal expression levels for gene i. For purpose of general discussion, the t_i values have been converted to z-values,

$$z_i = \Phi^{-1}(F_{100}(t_i)), \tag{1.1}$$

where Φ and F_{100} are the cumulative distribution functions (cdf) for standard normal and t_{100} distributions. Under the usual null assumption of i.i.d. normal sampling, z_i will have a standard $N(0,1)$ distribution – what we will call the

$$\textit{theoretical null } z_i \sim N(0,1) \tag{1.2}$$

The z_i's, null or non-null, are usually correlated in microarray situations but, fortunately, independence will not be required for the theory that follows.

Multiple comparisons has been an important topic for half a century. Rupert Miller's classic text "Simultaneous Statistical Inference", (1980), concerns doing two or three or maybe half a dozen tests at the same time. But 6033 simultaneous tests require a qualitative change in statistical theory: besides correcting p-values for multiplicities, the traditional concern of multiple comparisons methodology, an empirical Bayes kind of inter-case information forces itself upon frequentists and Bayesians alike.

Biostatistics is not the only beneficiary of computer-age scientific technology. Figure 2 shows the results of a California study comparing language skill differences between economically advantaged and disadvantaged high school students. An English-language proficiency test was administered in each of $N = 4138$ high schools, with say $n_{\text{adv}}(i)$ and $n_{\text{dis}}(i)$ being

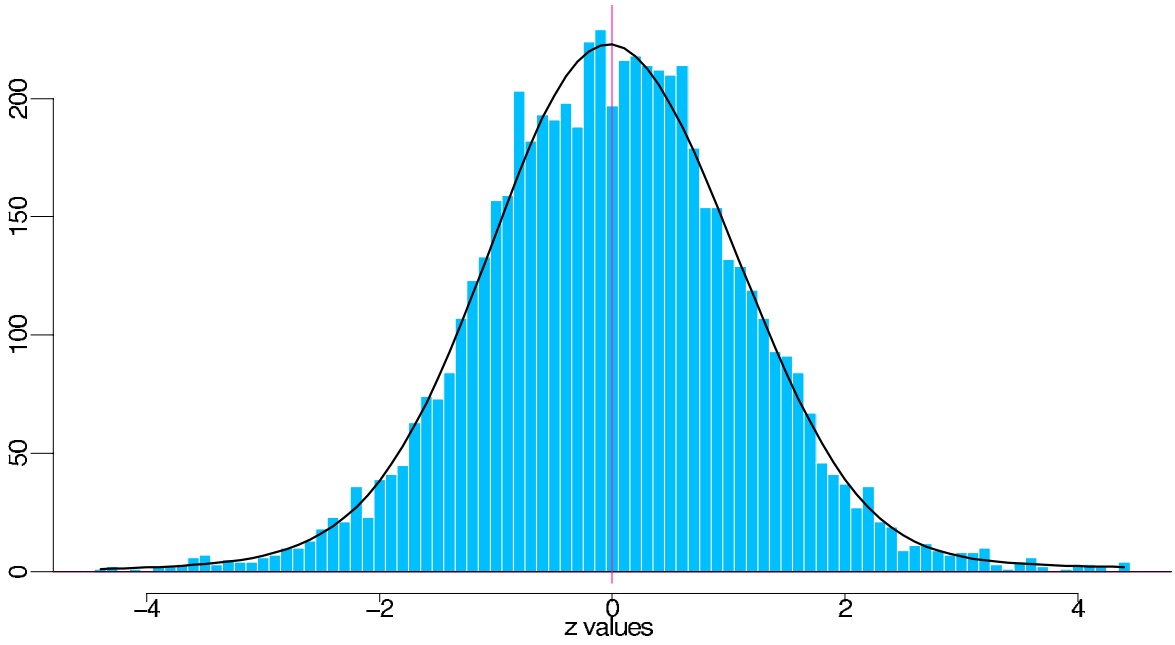

Figure 1: *6033 z-values from prostate cancer microarray study described in the text. Which of the 6033 genes show evidence of being non-null, i.e. expressed differently in prostate patients as compared to normal subjects? Solid curve is smooth fit to histogram heights. Data from Singh et al. (2002).*

the number of advantaged and disadvantaged students in the i^{th} school. A z-value for school i was computed according to

$$z_i = \frac{\widehat{p}_{\rm adv}(i) - \widehat{p}_{\rm dis}(i) - \Delta}{\left[\frac{\widehat{p}_{\rm adv}(1-\widehat{p}_{\rm adv}(i))}{n_{\rm adv}(i)} + \frac{\widehat{p}_{\rm dis}(i)(1-\widehat{p}_{\rm dis}(i))}{n_{\rm dis}(i)}\right]^{1/2}}, \tag{1.3}$$

where $\widehat{p}_{\rm adv}(i)$ and $\widehat{p}_{\rm dis}(i)$ were the observed test success rates, while the centering constant $\Delta = 0.229$ equaled median $(\widehat{p}_{\rm adv}(i))$ − median $(\widehat{p}_{\rm dis}(i))$.

A reasonable goal here would be to identify a subset of the high schools, presumably a small subset, in which the advantaged-disadvantaged performance differences are unusually large (in either direction). Large-scale hypothesis testing methods apply to the Education Data just as well as to the Prostate Study. However, the two data sets differ in a crucial way, relating to what is called the "empirical null" distribution in Figure 2. Section *5* discusses the required change in analysis strategy.

How should the statistician model, analyze, and report problems like the Prostate Study or the Education Data? This paper presents one point of view on the answer, my point of view as reported in a series of papers, Efron et al. (2001), Efron and Tibshirani (2002), Efron (2004, 2005, 2006abc) with 2006c being particularly relevant. The discussion here will be as nontechnical as possible, focusing on the statistical ideas involved and leaving mathematical details to the references. False discovery rates (Fdr), Benjamini and Hochberg's (1995) key contribution to large-scale testing, the organizing idea for what follows, begins the discussion in Section 2.

Statistical microarray analysis is a hot topic, and there are many other points of view in a rapidly growing literature, as nicely summarized in Dudoit et al. (2003). The aforementioned papers provide a range of references, some of which will be mentioned in the sequel.

2. False Discovery Rates Classic multiple comparisons theory concentrated attention on controlling the probability of a Type 1 error, a "false discovery", and much of the statistics microarray literature has pursued this same goal, as reviewed in Dudoit et al. (2003). However, with the number N of cases in the thousands, as in Figures 1 and 2, trying to limit the probability of even a single false discovery becomes unrealistically stringent. Benjamini and Hochberg's 1995 *False Discovery Rate* (Fdr) theory limits instead the expected proportion of false discoveries, a more relaxed criterion.

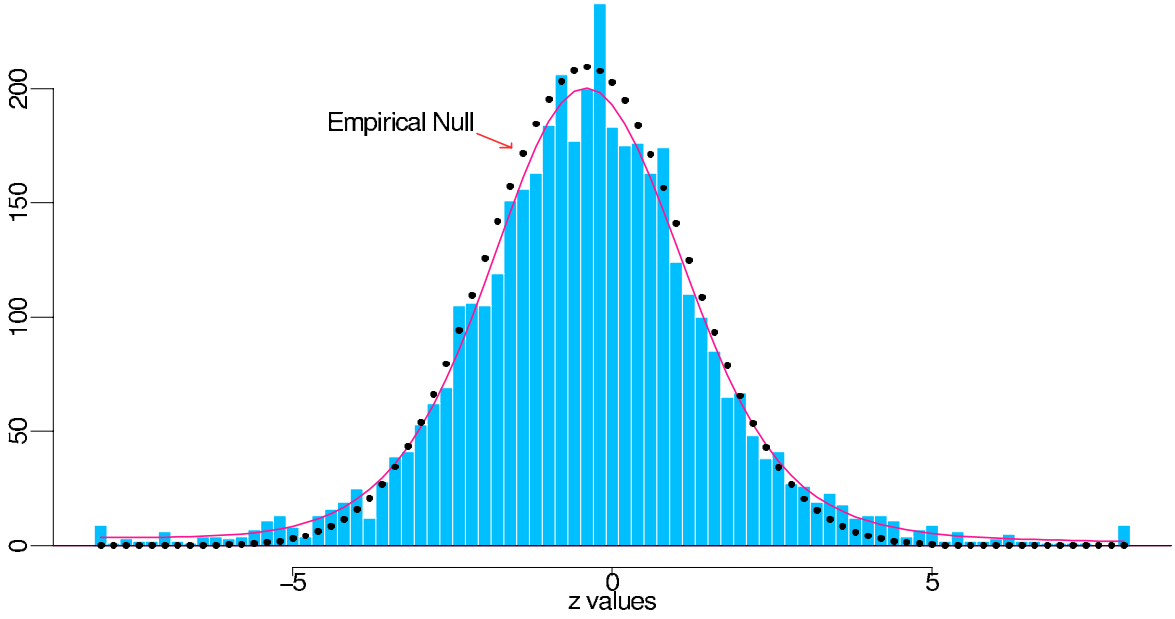

Figure 2: *Education Data; z-values (1.2) comparing pass rates for advantaged versus disadvantaged students, in* $N = 4183$ *California high schools. Solid Curve: smooth fit* $\widehat{f}$ *to histogram heights; Dots estimated empirical null* $\widehat{f}_0(z) \sim N(-.41, 1.51^2)$ *fit to central histogram heights, as discussed in Section 5. Data from Rogosa (2003).*

The *two-groups model* provides a simple framework for Fdr theory. We suppose that the N cases (genes for the Prostate Study, schools for the Education Data) are each either *null* or *non-null* with prior probability p_0 or $p_1 = 1 - p_0$, and with z-values having density either $f_0(z)$ or $f_1(z)$,

$$\begin{array}{ll} p_0 = Pr\{\text{null}\} & f_0(z) \text{ density if null} \\ p_1 = Pr\{\text{non-null}\} & f_1(z) \text{ density if non-null,} \end{array} \tag{2.1}$$

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 Figure 1 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 non-null genes to no more than 10%. Model (2.1) has been widely used, as in Lee et al. (2000), Newton et al. (2001), and Efron et al. (2001), not always with restriction (2.2).

False discovery rate 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

$$\text{Fdr}(z) \equiv Pr\{\text{null}|Z \leq z\} = p_0F_0(z)/F(z). \tag{2.3}$$

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

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

(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 non-null 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),

$$f(z) = p_0 f_0(z) + p_1 f_1(z), \tag{2.6}$$

Bayes rule gives

$$\text{fdr}(z) \equiv Pr\{\text{null}|Z = z\} = p_0 f_0(z)/f(z) \tag{2.7}$$

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

There is a simple relationship between $\text{Fdr}(z)$ and $\text{fdr}(z)$,

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

"E_f" indicating expectation with respect to the mixture density $f(z)$. That is, $\text{Fdr}(z)$ is the mixture average of $\text{fdr}(Z)$ for $Z \le z$. In the usual situation where $\text{fdr}(z)$ decreases as $|z|$ gets large, $\text{Fdr}(z)$ will be smaller than $\text{fdr}(z)$. Intuitively, if we decide to label all genes with z_i less than some negative value z_0 as "non-null", then $\text{fdr}(z_0)$, the false discovery rate at the boundary point z_0, will be greater than $\text{Fdr}(z_0)$, the average false discovery rate beyond the boundary.

For Lehmann alternatives

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

it turns out that

$$\log\left\{\frac{\text{fdr}(z)}{1-\text{fdr}(z)}\right\} = \log\left\{\frac{\text{Fdr}(z)}{1-\text{Fdr}(z)}\right\} + \log\left(\frac{1}{\gamma}\right), \tag{2.10}$$

so

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

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

The heavy curve in Figure 3 is an estimate of the local false discovery rate $\text{fdr}(z)$ for the Prostate Study, (2.7) based on the algorithm *locfdr* discussed in Section 3. For $|z_i| \le 2.0$ the curve is near 1, so we would definitely not report such genes as interesting possibilities since they are almost certainly null cases. Things get more interesting as z_i gets farther away from zero. 51 of the 6033 genes, 26 on the right and 25 on the left, have $\text{fdr}(z_i) \le 0.20$, a conventional reporting point motivated in Efron (2006c). We could report this list of 51 to the researchers as good bets (better than 80%) for being genuinely non-null cases. (By comparison, a standard Benjamini-Hochberg procedure, (2.5) with $q = 0.1$, reports 60 non-null genes, 28 on the right and 32 on the left.)

The beaded curves in Figure 3 are smoothed versions of (2.4), estimates of $\text{Fdr}(z)$, (2.3), and the corresponding right tail quantity

$$Pr\{\text{null}|Z \ge z\} = p_0[1 - F_0(z)]/[1 - F(z)]. \tag{2.12}$$

At the points where $\widehat{\text{fdr}}(z) = 0.2$, the Fdr estimates are 0.108 on the left and .081 on the right, corresponding roughly to $\gamma = 1/2$ in (2.9).

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 $\text{fdr}_i(z) = Pr\{\text{gene}_i \text{ null}|z_i = z\}$:

$$\text{fdr}_i(z) = \text{fdr}(z)\frac{r_i}{1 - (1 - r_i)\text{fdr}(z)} \qquad \left[r_i = \frac{p_0(i)}{1 - p_0(i)}\Big/\frac{p_0}{1 - p_0}\right], \tag{2.13}$$

where $\text{fdr}(z) = p_0 f_0(z)/f(z)$ 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 $\text{fdr}_i(z_i) = 0.069$.

3. Estimating Fdr and fdr Bayesian model (2.1), the two-groups model, might provoke the usual frequentist criticism: it assumes the statistician knows quantities that in most practical situations will be obscure. However, a marvelous thing happens in problems where thousands of similar situations are observed simultaneously: we can use all of the data to *estimate* Bayes rule, usually in a frequentist way, and then proceed as Bayesians. This is the *empirical Bayes* approach pioneered by Robbins and Stein in the 1950's, Efron (2003), now come to fruition in the Twenty First Century.

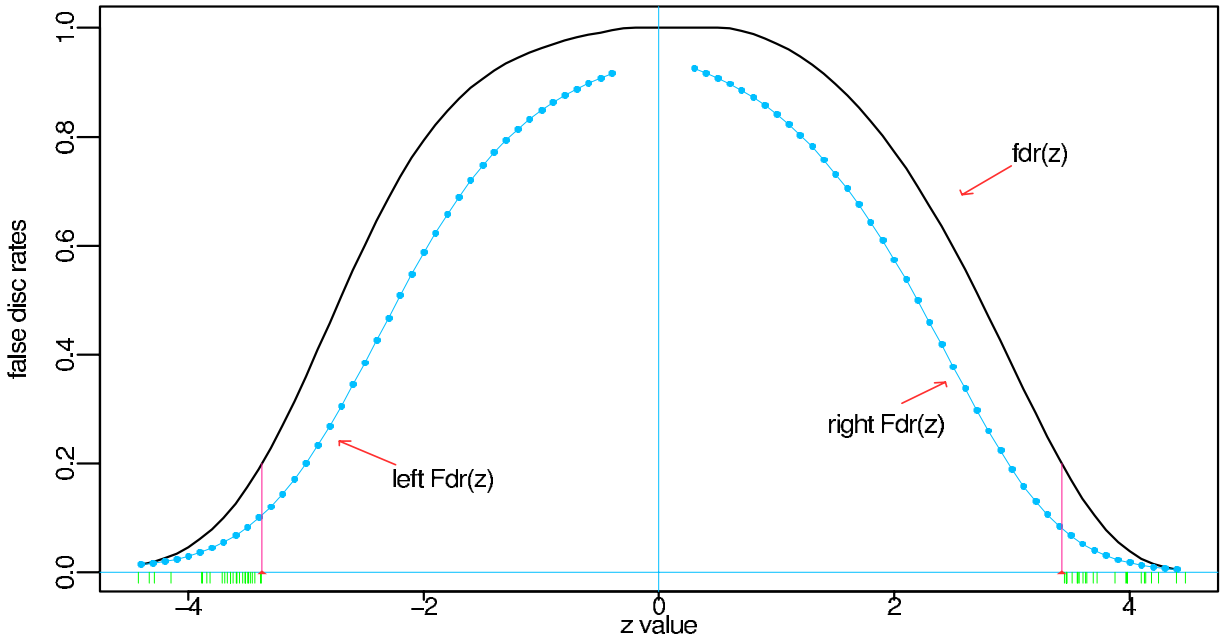

Figure 3: *Heavy curve is estimated local false discovery rate fdr(z), (2.7), for Prostate Data, using locfdr methodology of Section 3. Beaded curves are estimates of tail areas false discovery rates Fdr(z) as for (2.3) and the corresponding right tail Fdr. 51 genes, 26 on right and 25 on left, indicated by dashes, have* $fdr(z_i) \leq 0.2$.

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 Figure 1, where it is easy to believe in the theoretical null distribution (1.2),

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

If we had only gene i's data to test, classic theory would tell us to compare z_i with $f_0(z) = \varphi(z)$, rejecting the null hypothesis of no difference between normal and prostate cancer expression levels if $|z_i|$ exceeded 1.96.

For the moment I will take p_0, the prior probability of a gene being null, as known. Section 5 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

$$y_k = \#\{z_i \quad \text{in} \quad bin\ k\},\ k = 1, 2, \ldots K. \tag{3.2}$$

The prostate data histogram in Figure 1 has $K = 89$ bins of width $\Delta = 0.1$.

We take the y_k to be independent Poisson counts,

$$y_k \overset{\text{ind}}{\sim} Po(\nu_k) \qquad k = 1, 2, \ldots K, \tag{3.3}$$

with the unknown ν_k proportional to density $f(z)$ at midpoint "x_k" of the k^{th} bin, approximately

$$\nu_k = N\Delta f(x_k). \tag{3.4}$$

Modeling $\log(\nu_k)$ as a p^{th} degree polynomial function of x_k makes (3.3)-(3.4) a standard Poisson general linear model (GLM). The choice $p = 7$, used in Figures 1 and 2, amounts to estimating $f(z)$ by maximum likelihood within the seven-parameter exponential family

$$f(z) = \exp\left\{\sum_{j=0}^{7} \beta_j z^j\right\}. \tag{3.5}$$

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*, 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 (3.5).

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

$$\widehat{\text{fdr}}(z) = p_0 f_0(z)/\widehat{f}(z), \tag{3.6}$$

with $\widehat{f}(z)$ constructed as above, $f_0(z) = \varphi(z)$ in (3.1), and $p_0 = 0.94$ as estimated in Section 5.

At this point the reader might notice an anomaly: if $p_0 = 0.94$ of the $N = 6033$ genes are null, then about $(1 - p_0) \cdot 6033 = 362$ should be non-null, but only 51 were reported in Section 2. The trouble is that most of the non-null 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, as discussed in the next Section.

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

$$\overline{\text{fdr}} = 2.68/17 = 0.16 \tag{3.7}$$

as an estimated false discovery rate. (The smoothed estimate in Figure 3 is $\widehat{\text{fdr}} = 0.24$.) 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 (3.7) catches the main idea.

Notice that we do not need the null genes to all 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 *expectations*, not *probabilities*. The non-null 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 (3.7): 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 (2.13).

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 (3.6) will efficiently estimate $\text{fdr}(z)$. Independence of the z_i's is *not* required, only that $\widehat{f}(z)$ is a reasonably close estimate of $f(z)$.

z	fdr	local	(formula)	tail
1.5	.88	**.05**	(.05)	**.05**
2.0	.69	**.08**	(.09)	**.05**
2.5	.38	**.09**	(.10)	**.05**
3.0	.12	**.08**	(.10)	**.06**
3.5	.03	**.10**	(.13)	**.07**
4.0	.005	**.11**	(.15)	**.10**

Table 1: *Boldface* standard errors of $\log \widehat{\text{fdr}}(z)$, (local fdr), and $\log \widehat{\text{Fdr}}(z)$, (tail area Fdr); 250 replications of model (3.8), with $N = 1500$ cases per replication; "fdr" is true value (2.7). *Parentheses* show average from formula (5.9), Efron (2006b).

Table 1 reports on a small simulation study in which

$$z_i \overset{\text{ind}}{\sim} N(\mu_i, 1) \begin{cases} \mu_i = 0 \text{ with probability } 0.9 \\ \mu_i \sim N(3,1) \text{ with probability } 0.1 \end{cases} \tag{3.8}$$

for $i = 1, 2, \ldots, N = 1500$. The table shows standard deviations for $\log(\widehat{\text{fdr}}(z))$, (3.6), from 250 simulations of (3.11), and also using a delta-method formula derived in Section 5 of Efron (2006b), incorporated in the *locfdr* algorithm. Rather than (3.5), $f(z)$ was modeled by a seven-parameter natural spline basis, *locfdr's* default, though this gave nearly the same results as (3.5). 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 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 $N^{-\frac{1}{2}}$, 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 schizophrenic methodology, with alternating episodes of frequentist and Bayesian activity. Frequentists may prefer $\widehat{\text{Fdr}}$ (or $\overline{Fdr}$, (2.5)) to $\widehat{\text{fdr}}$ because of connections with classical tail-area hypothesis testing, or because cdfs 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, semi-parametric, nonparametric, and Bayes methods; Pan et al. (2003), Pounds and Morris (2003), Allison et al. (2004), Heller and Qing (2003), Broberg (2005), Aubert et al. (2004), Liao et al. (2004), and Do et al. (2004), 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 that provided the formula values in Table 1.

All of this assumes that the theoretical null distribution (3.1) is in fact correct for the problem at hand. By no means is this always a safe assumption! Section 5 discusses situations like the Education Data of Figure 2, where the theoretical null is grossly incorrect; and where the null distribution itself must be estimated from the data. Estimation efficiency becomes a more serious problem in such *empirical null* situations.

4. Power Calculations and Non-Null Counts Most of the statistical microarray literature has concentrated on controlling Type I error, the false rejection of genuinely null cases. Power, the probability of correctly rejecting non-null cases, deserve attention too. In some ways, power calculations are more accessible in large-scale testing situations than in individual problems, with a single data set like the Prostate Study being able to provide its own power diagnostics. This section shows the local false discovery estimate $\widehat{\text{fdr}}(z)$, (3.6), in action as a diagnostic tool.

The histogram in Figure 1 has 89 bins of width $\Delta = 0.1$ each, spanning the range $-4.45 \le z \le 4.45$. The histogram bars are of height "Y_k",

$$y_k = \#\{z_i \text{ in bin } k\}, \quad k = 1, 2, \ldots K = 89. \tag{4.1}$$

"Wouldn't it be great", one might say, "if we could see the histogram of z-values for just the non-null cases, the ones we are interested in?" In fact, we *can* estimate the non-null histogram from $\widehat{\text{fdr}}(z)$.

The vertical bars in Figure 4 are of height

$$Y_k = [1 - \widehat{\text{fdr}}_k] \cdot y_k, \tag{4.2}$$

where $\widehat{\text{fdr}}_k$ is $\widehat{\text{fdr}}(z)$ evaluated at the centerpoint of bin k. Since $1 - \widehat{\text{fdr}}_k$ approximates the non-null probability for a gene in bin k, Y_k is an obvious estimate for the expected number of non-null genes. The total non-null count ΣY_k equals about $(1 - p_0) \cdot N$, 362 in this case, resolving the "anomaly" noted after (3.6) in Section 3.

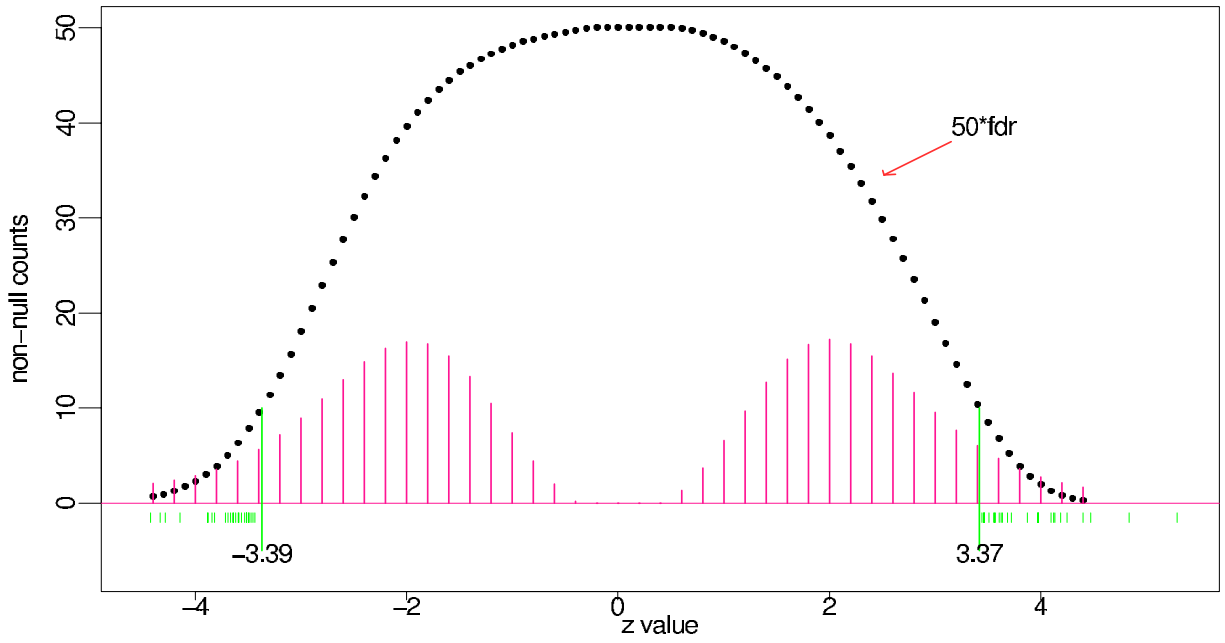

Figure 4: *Vertical bars are non-null counts for Prostate data, estimates of the histogram of z-values for only the non-null cases. Dotted curve is* $50 \cdot \widehat{\text{fdr}}(z)$*, estimated local false discovery rate from Figure 3. Power diagnostics obtained by comparing non-null histogram with* $\widehat{\text{fdr}}$ *curve. For example the expected value of* $\widehat{\text{fdr}}$ *with respect to non-null histogram is 0.68, indicating low power for Prostate study.*

Power diagnostics are obtained from comparisons of $\widehat{\text{fdr}}(z)$ with the non-null histogram. High power would be indicated if $\widehat{\text{fdr}}_k$ was small where Y_k was large. That obviously is not the case in Figure 4. A simple power diagnostic is

$$\widehat{E\ \text{fdr}}_1 = \sum_{k=1}^{K} Y_k\ \widehat{\text{fdr}}_k \Big/ \sum_{k=1}^{K} Y_k, \tag{4.3}$$

the expected non-null fdr. We want $\widehat{E\ \text{fdr}}_1$ to be small, perhaps near 0.2, so that a typical non-null gene will show up on a list of likely prospects. The Prostate data has $\widehat{E\ \text{fdr}}^{(1)} = 0.68$, indicating low power. If the whole study were rerun we might expect a different list of 51 likely non-null genes, barely overlapping with the first list.

Going further, we can examine the entire non-null histogram of $\widehat{\text{fdr}}(z)$ rather than just its expectation. The non-null cdf of $\widehat{\text{fdr}}$ is estimated by

$$\widehat{G}(t) = \sum_{k:\widehat{\text{fdr}}_k \le t} Y_k \Big/ \sum_{k} Y_k \tag{4.4}$$

Figure 5 shows $\widehat{G}(t)$ for the Prostate study and for the first of the 250 simulations from model (3.8) in Table 1. The figure suggests poor power for the Prostate Study, with only 11% probability that a non-null gene will have its estimated $\text{fdr}(z)$ values less than 0.2; conversely, model (3.8) is a high-power situation. Section 3 of Efron (2006b) discusses more elaborate power diagnostics.

Graphs such as Figure 5 help answer the researcher's painful question "Why aren't the cases we expected on your list of likely non-null outcomes?" Low power is often the reason. For the Prostate data, *most* of the non-null genes will not turn up on the list of low fdr cases. The *R* program *locfdr*, used to construct Figure 3, also returns $\widehat{\text{Efdr}}_1$ and $\widehat{G}(t)$.

5. Empirical Estimation of the Null Distribution Classical hypothesis testing assumes exact knowledge of the null density $f_0(z)$, as in (1.2). Benjamini and Hochberg's Fdr controlling algorithm (2.5) is based on the same assumption, as is the local false discovery estimate $\widehat{\text{fdr}}(z) = p_0 f_0(z)/\widehat{f}(z)$ in (3.6). In fact almost all of the microarray statistics literature begins with the assumption that $f_0(z)$, the null density in (2.1), is known on theoretical grounds, or can be recovered from permutation calculations (which usually produce only minor corrections, and discussed later).

Use of the theoretical null is mandatory in classic one-at-a-time testing where theory provides the only information on null behavior. But large-scale simultaneous testing differs

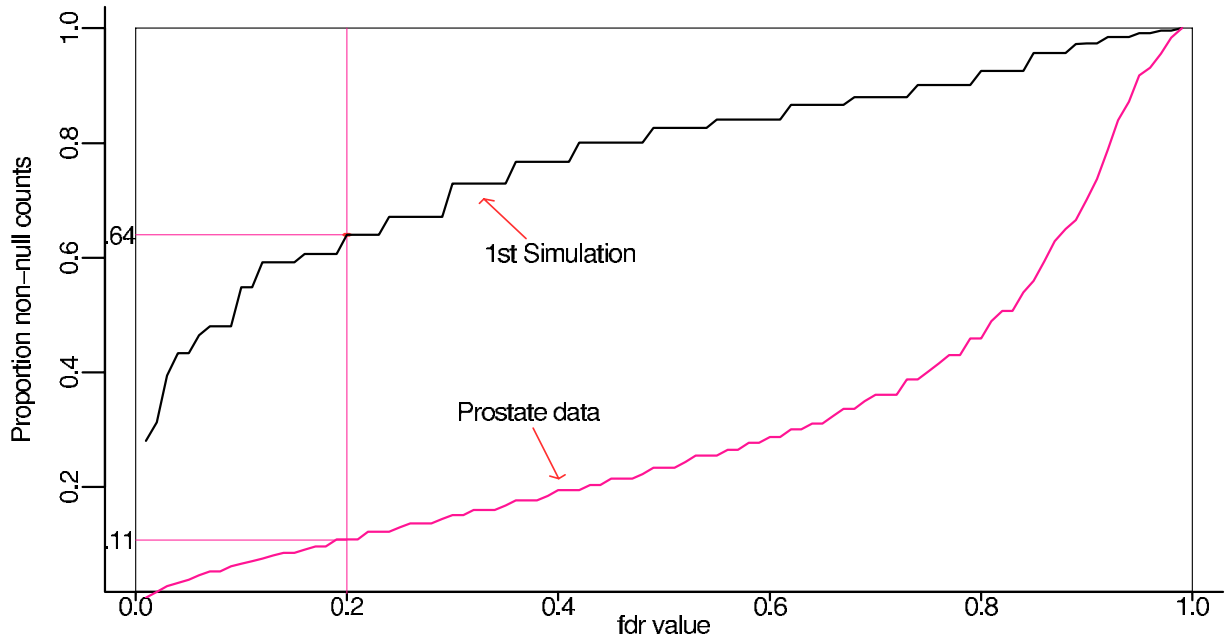

Figure 5: *Estimated non-null cdf for fdr, (4.4). Prostate data has low power, with only 11% of the non-null cases having fdr values* ≤ 0.20*, compared to 64% first simulation of (3.8).*

in two important ways: the theoretical null may appear obviously suspect, as it is in Figure 2 where the center of the z-value histogram is more than 50% wider than $z \sim N(0,1)$ would predict; and it may be possible to estimate the null distribution itself from the histogram.

If this last statement sounds heretical, the basic idea is simple enough: we make the "zero assumption",

$$\textit{Zero assumption} \quad \text{most of the z-values near 0 come from null genes,} \tag{5.1}$$

(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 6 shows the geometric method in action on the Education data. The heavy solid curve is $\log \widehat{f}(z)$, fit from (3.5) using Lindsey's method. The two groups model (2.1) 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

$$\varphi_{\delta_0,\sigma_0}(z) \equiv (2\pi\sigma_0^2)^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2}\left(\frac{z-\delta_0}{\sigma_0}\right)^2 \right\}, \tag{5.2}$$

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

$$\log f(z) \doteq \log p_0 - \frac{1}{2}\left\{\frac{\delta_0^2}{\sigma_0^2} + \log(2\pi\sigma_0^2)\right\} + \frac{\delta_0}{\sigma_0^2}z - \frac{1}{2\sigma_0^2}\, z^2. \tag{5.3}$$

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 (5.3) yields estimates $(\widehat{\delta}_0, \widehat{\sigma}_0, \widehat{p}_0)$, for instance $\widehat{\sigma}_0 = (2\widehat{\beta}_2)^{-\frac{1}{2}}$,

$$\widehat{\delta}_0 = -0.41, \quad \widehat{\sigma}_0 = 1.51, \quad \text{and } \widehat{p}_0 = 0.92 \tag{5.4}$$

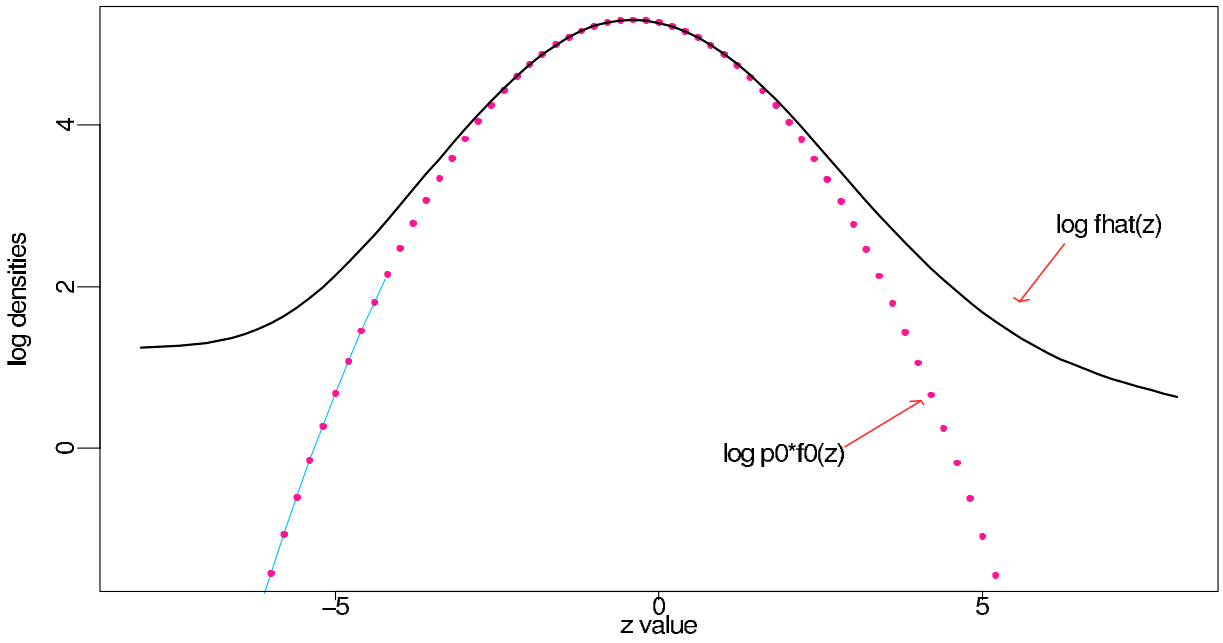

Figure 6: *Geometric estimation of null proportion p_0 and empirical null mean and standard deviation (δ_0, σ_0) for the Education Data of Figure 2. Heavy curve is $\log \widehat{f}(z)$, estimated as in (3.2)-(3.5); beaded curve is best quadratic approximation to $\log \widehat{f}(z)$ around its maximum.*

for the Education Data. Trying the same method with the theoretical null, that is taking $(\delta_0, \sigma_0) = (0, 1)$, gives a very poor fit to $\log \widehat{f}(z)$, miscentered and much too narrow. Figure 7 shows the comparison in terms of the densities themselves rather than their logs.

The analytic method makes more explicit use of the zero assumption, stipulating that the non-null 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

$$P_0(\delta_0, \sigma_0) = \Phi\left(\frac{b - \delta_0}{\sigma_0}\right) - \Phi\left(\frac{a - \delta_0}{\sigma_0}\right) \quad \text{and} \quad \theta = p_0 P_0. \tag{5.5}$$

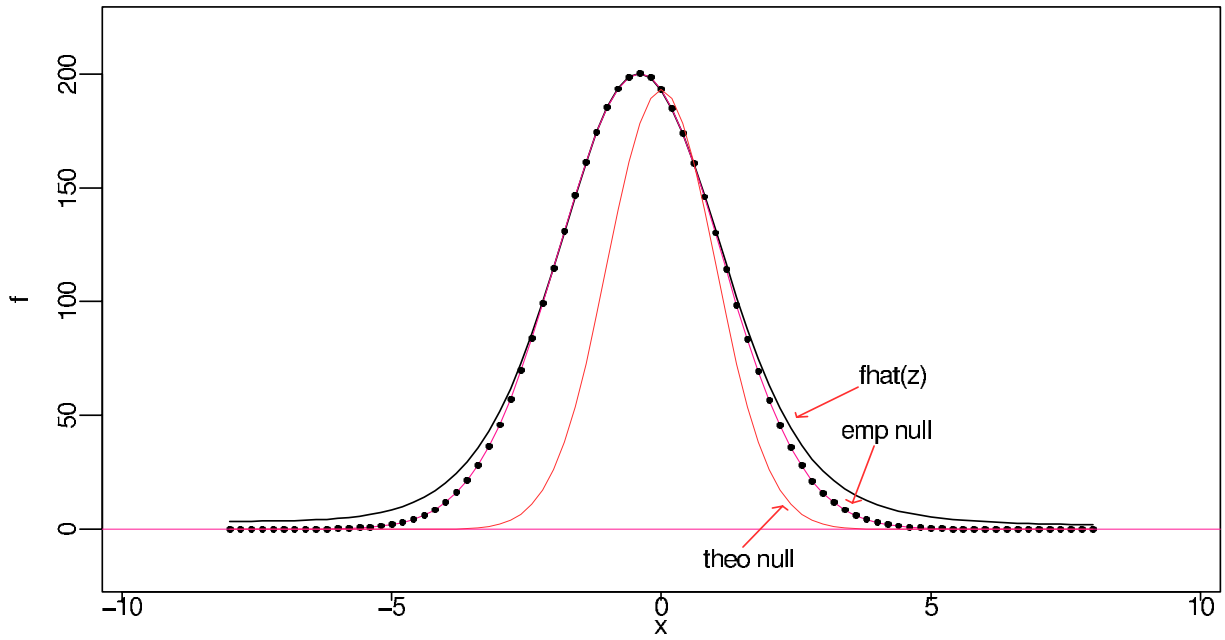

Figure 7: *Solid curve is $\widehat{f}(z)$, (3.2)-(3.5), fit to Education Data; beaded curve is $\widehat{p}_0\widehat{f}_0(z)$, empirical null fitted geometrically as in Figure 6. Light curve is estimate of $p_0 f_0(z)$ based on theoretical null (1.2).*

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

$$f_{\delta_0,\sigma_0,p_0}(\mathbf{z}_0) = [\theta^{N_0}(1-\theta)^{N-N_0}]\left[\prod_{z_i\in\mathbf{z}_0}\frac{\varphi_{\delta_0,\sigma_0}(z_i)}{P_0(\delta_0,\sigma_0)}\right]. \tag{5.6}$$

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 (-0.41, 1.58, 0.94) for the Education 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 if $p_0 \geq 0.90$. 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. $p_i = 2(1 - F_{100}(|t_i|))$ rather than (1.1), 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

	Geometric			Analytic		
	mean	**stdev**	(formula)	mean	**stdev**	(formula)
$\widehat{\delta}_0$:	0.02	**.056**	(.062)	0.04	**.031**	(.032)
$\widehat{\sigma}_0$:	1.02	**.029**	(.033)	1.04	**.031**	(.031)
$\widehat{p}_0$:	0.92	**.013**	(.015)	0.93	**.009**	(.011)

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 Efron (2006b), as implemented in *locfdr*.

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 of Efron (2006c) 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 Table 3 with Table 1 reveals factors of two or three 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 area Fdr's rather than local fdr's does not help – here the local version is less variable.)

Then why not just always use the theoretical null for $f_0(z)$? The answer is that in situations like Figure 2, there is direct evidence that something may be wrong with the null assumption $z \sim N(0, 1)$. Large-scale simultaneous testing has as its usual purpose the selection of a *small* subset of cases worthy of further consideration, but for the Education Data it is easy to show that the proportion of non-null cases p_1 must exceed 41% if we insist that $f_0(z) = N(0, 1)$. Saying that 41% of the high schools are unusual really says nothing at all.

Section 5 of Efron (2006c) lists several reasons why the theoretical null distribution might fail in practice, among which one seems particularly dangerous here: *unobserved covariates* in the high schools, class size, economic status, racial composition etc., tend to expand the z-value histogram, null and non-null cases alike. Using an empirical null compensates for all such effects, at the expense of increased estimation variability. See Section 4 of Efron (2004)

		EMPIRICAL NULL		
z	fdr	local	(formula)	tail
1.5	.88	**.04**	(.04)	**.10**
2.0	.69	**.09**	(.10)	**.15**
2.5	.38	**.16**	(.16)	**.23**
3.0	.12	**.25**	(.25)	**.32**
3.5	.03	**.38**	(.38)	**.42**
4.0	.005	**.50**	(.51)	**.52**

Table 3: Standard errors of $\log\widehat{\text{fdr}}(z)$ and $\log\widehat{\text{Fdr}}(z)$ as in Table 1, but now using empirical nulls (geometric method). Delta-method formula estimates for $sd\{\log\widehat{\text{fdr}}(z)\}$ are included in output of *R* program *locfdr*.

for an example and discussion. Section 5 of Efron (2006c) presents two microarray examples, one where the theoretical null is much too narrow, and another where it is much too wide.

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 Education 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).

Permutation methods play a major role in the microarray statistics literature. For the prostate data one can randomly permute the 102 microarrays (i.e. permute the columns of the data matrix X), recalculate the t statistics between the first 50 and last 52 columns, and convert to z-values as in (1.1). Doing this say 1000 times produces a "permutation null" distribution, almost perfectly $N(0,1)$ for the Prostate data. Unfortunately, this tends to happen whether or not there are unobserved covariates, or other of the problems that undermine the theoretical null distribution.

Section 5 of Efron (2006c) discusses both the strengths and limitations of permutation testing for large-scale inference. Permutation methods are definitely useful, but they cannot substitute for empirical null calculations in most settings.

6. Summary In a sense, statistical theory has been living off of intellectual capital accumulated in the first half of the Twentieth Century, capital that may now seem rather spent when facing situations like those in Figures 1 and 2. The Fisher-Neyman-Pearson theory of hypothesis testing was fashioned for a scientific world where experimentation was slow and difficult, producing small data sets intended to answer single questions. It has been wonderfully successful within this milieu, combining elegant mathematics and limited computational equipment to produce scientificallly dependable answers in a variety of application areas.

"Drinking from a fire-hose" is the way one of my colleagues described the influx of data from a typical microarray experiment. Here I have tried to indicate one approach to handling the fire-hose problem as it concerns simultaneous hypothesis testing. Massive data sets like those in Section 1 are misleadingly comforting in their suggestion of great statistical accuracy. The size and power calculations of Sections 3-5, carried out via the *locfdr* algorithm (available through CRAN) show that the ability to detect specific interesting cases may still be quite low. Efficient statistical inference is still a necessity, even with the largest data sets.

As I said to begin with, these are exciting times for statisticians, for applied statisticians certainly, but for theorists too. What I called "empirical Bayes information" accumulates in a way that is not well understood yet, but seems to me to play a central role in large-scale simultaneous inference. We are living in an age of heroic statistical methodology – my hope is that an heroic theory to justify it cannot be far behind.

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", *Computational Statistics and Data Analysis* **39**, 1-20.

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

Benjamini, Y. and Yekutieli, D. (2001). "The control of the false discovery rate under dependency", *Ann. Stat.* **29**, 1165-88.

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

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

Efron, B. and Tibshirani, R. (1996). "Using specially designed exponential families for density estimation", *Annals Stat.* **24**, 2431-61.

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

Efron, B., and Tibshirani, R. (2002). "Empirical Bayes methods and false discovery rates for microarrays", *Genetic Epidemiology* **23**, 70-86.

Efron, B. (2004). "Large-scale simultaneous hypothesis testing: the choice of a null hypothesis", *JASA* **99**, 96-104.

Efron, B. (2005). "Local false discovery rates", http://www-stat.stanford.edu/∼brad/papers/False.pdf

Efron, B. (2006a). "Size, power, and false discovery rates" (to appear, *Annals of Statistics*) http://www-stat.stanford.edu/∼brad/papers/Size.pdf

Efron, B. (2006b). "Correlation and large-scale simultaneous significance testing", http://www-stat.stanford.edu/∼brad/papers/Correlation-2006.pdf (to appear *JASA*).

Efron, B. (2006c). "Microarrays, empirical Bayes, and the two-groups model". (To appear *Statistical Science*).

Heller, G. and Qing, J. (2003). "A mixture model approach for finding informative genes in microarray studies", Unpublished.

Langass, M., Lindquist, B. and Ferkingstad, E. (2005). "Estimating the proportion of true null hypotheses, with application to DNA microarray data", *JRSS-B* **67**, 555-72.

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. Nat. Acad. Sci.* **97**, 9834-38.

Lehmann, E. and Romano, J. (2005). "Generalizations of the Familywise Error Rate", *Annals Stat.* **33** , 1138-1154.

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.

Miller, R. (1980). *Simultaneous Statistical Inference*, Springer, New York.

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, *Jour. Comp. Biology* **8**, 37-52.

Pan, W., Lin, J., Le, C. (2003). "A mixture model approach to detecting differentially expressed genes with microarray data", *Functional & Integrative Genomics* **3**, 117-24.

Pawitan, Y., Murthy, K., Michiels, S. and Ploner, A. (2005). "Bias in the estimation of false discovery rate in microarray studies", *Bioinformatics* **21**, 3865-72.

Pounds, S. and Morris, S. (2003). "Estimating the occurrence of false positions and false negatives in microarray studies by approximating and partitioning the empirical distribution of the p-values", *Bioinformatics* **19**, 1236-42.

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

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*, 1:302-209.

Storey, J. (2002). "A direct approach to false discovery rates", *Journal of the Royal Statistical Society*, Ser. B, **64**, 479-498.