Large-Scale Simultaneous Hypothesis Testing: The Choice of a Null Hypothesis

Bradley EFRON

Current scientific techniques in genomics and image processing routinely produce hypothesis testing problems with hundreds or thousands of cases to consider simultaneously. This poses new difficulties for the statistician, but also opens new opportunities. In particular, it allows empirical estimation of an appropriate null hypothesis. The empirical null may be considerably more dispersed than the usual theoretical null distribution that would be used for any one case considered separately. An empirical Bayes analysis plan for this situation is developed, using a local version of the false discovery rate to examine the inference issues. Two genomics problems are used as examples to show the importance of correctly choosing the null hypothesis.

KEY WORDS: Empirical Bayes; Empirical null hypothesis; Local false discovery rate; Microarray analysis; Unobserved covariates.

1. INTRODUCTION

Until recently, "simultaneous inference" meant considering two or five or perhaps even 10 hypothesis tests at the same time, as in Miller's classic text (Miller 1981). Rapid progress in technology, particularly in genomics and imaging, has vastly upped the ante for simultaneous inference problems. Now 500 or 5,000 or even 50,000 tests may need to be evaluated simultaneously, raising new problems for the statistician, but also opening new analytic opportunities. This article explores choosing an appropriate null hypothesis in large-scale testing situations, and how this choice affects well-known inference methods, such as the false discovery rate (FDR).

Simultaneous hypothesis testing begins with a collection of null hypotheses,

$$H_1, H_2, \ldots, H_N; \tag{1}$$

corresponding test statistics, possibly not independent,

$$Y_1, Y_2, \ldots, Y_N; \tag{2}$$

and their p values, $P_1, P_2, \ldots, P_N$, with P_i measuring how strongly y_i, the observed value of Y_i, contradicts H_i; for instance, $P_i = \Pr_{H_i}\{|Y_i| > |y_i|\}$. "Large-scale" means that N is a big number, say at least $N > 100$.

It is convenient, although not necessary, to work with *z-values* instead of the Y_i's or P_i's,

$$z_i = \Phi^{-1}(P_i), \qquad i = 1, 2, \ldots, N, \tag{3}$$

with Φ indicating the standard normal *cumulative distribution function* (cdf), for example, $\Phi^{-1}(.95) = 1.645$. If H_i is exactly true, then z_i will have a standard normal distribution,

$$z_i | H_i \sim \mathrm{N}(0, 1). \tag{4}$$

I call (4) the *theoretical null hypothesis*.

Our motivating example concerns a study of 1,391 patients with human immunodeficiency virus (HIV) infection, investigating which of 6 protease inhibitor (PI) drugs cause mutations at which of 74 sites on the viral genome. Each patient provided a vector of predictors,

$$\mathbf{x} = (x_1, x_2, \ldots, x_6), \tag{5}$$

with $x_j = 1$ or 0 indicating whether or not the patient used PI_j, $1 \le \sum_1^6 x_j \le 6$; and a vector of responses,

$$\mathbf{v} = (v_1, v_2, \ldots, v_{74}), \tag{6}$$

$v_k = 1$ or 0 indicating whether or not a mutation occurred at site k. Remark A of Section 7 describes the study in more detail.

For each of the 74 genomic sites, a separate logistic regression analysis was run using all 1,391 cases, with that site's mutation indicators as responses and the PI indicators as predictors. Together these yielded $444 = 6 \times 74$ z-values, one for testing each null hypothesis that drug j does not cause mutations at site k, $j = 1, 2, \ldots, 6$ and $k = 1, 2, \ldots, 74$. The z-values were based on the usual approximation

$$z_i = y_i / se_i, \qquad i = 1, 2, \ldots, 444, \tag{7}$$

[using a single subscript i in place of (j, k)] where y_i is the maximum likelihood estimate (MLE) of the logistic regression coefficient and se_i is its approximate large-sample standard error.

Figure 1 shows a histogram of the 444 z-values, with negative z_i's indicating greater mutational effects. The smooth curve, $f(z)$, is a natural spline with 7 df, fit to the histogram counts by Poisson regression. It emphasizes the *central peak* near $z = 0$, presumably the large majority of uninteresting drug–site combinations that have negligible mutation effects. Near its center, the peak is well described by a normal density with mean $-.35$ and standard deviation 1.20, which will be called the *empirical null hypothesis*,

$$z_i | H_i \sim \mathrm{N}(-.35, 1.20^2). \tag{8}$$

Section 3 describes the estimation methodology for (8), with a brief discussion of the normality assumption in Remark D of Section 7.

The difference between the theoretical null N(0, 1) and empirical null $\mathrm{N}(-.35, 1.20^2)$ may not seem worrisome here, but it will be shown that it substantially affects any simultaneous inference procedure. More dramatic example is given in Section 6, for a microarray analysis in which going from the theoretical to empirical null totally negates any findings of significance. Situations going in the reverse direction can also occur.

Bradley Efron is Professor, Department of Statistics, Stanford University, Stanford, CA 94305 (E-mail: *brad@stat.stanford.edu*). The author thanks Robert Shafer, David Katzenstein, and Rami Kantor for bringing the drug mutation data to his attention, and Robert Tibshirani for several helpful discussions.

Journal of the American Statistical Association
March 2004, Vol. 99, No. 465, Theory and Methods
DOI 10.1198/016214504000000089

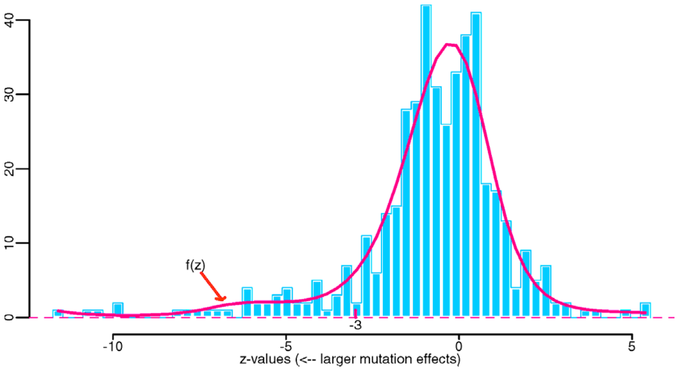

Figure 1. Histogram of 444 z-Values From the Drug Mutation Analysis. The smooth curve f(z) is a natural spline fit to histogram counts. The central peak near $z = 0$ is approximately $N(-.35, 1.20^2)$, the "empirical null hypothesis." Simultaneous hypothesis tests for the 444 cases depend critically on the choice between the empirical or theoretical N(0, 1) null.

In classic situations involving only a single hypothesis test, one must, out of necessity, use the theoretical null hypothesis, $z \sim \mathrm{N}(0, 1)$. The main point of this article is that large-scale testing situations permit empirical estimation of the null distribution. Sections 3–5 explore reasons why the empirical and theoretical null might differ, and which might be preferable in different situations.

There are scientific as well as statistical differences between small-scale and large-scale hypothesis testing situations. A single hypothesis test is most often run with the expectation and hope of rejecting the null, "with 80% power" in a typical clinical trial. Nobody wants to reject 80% of $N = 5{,}000$ null hypotheses. The usual point of large-scale testing is to identify a small percentage of interesting cases that deserve further investigation. Although we are not exactly looking for a needle in a haystack, we do not want the whole haystack either. An important assumption of what follows is that the proportion of interesting cases is small, perhaps 1% or 5% of N, but not more than 10%. This is made explicit in Section 2, in the description of the local false discovery rate as an analytic tool for large-scale testing. There are situations in which the 10% limit is irrelevant (e.g., in constructing prediction models), but these lie outside our purpose here.

The terminology "Interesting/Uninteresting" used in this article in preference to "Significant/Nonsignificant" is discussed near the end of Section 5. We conclude in Sections 7 and 8 with remarks, including most of the technical details, and a summary.

2. THE LOCAL FALSE DISCOVERY RATE

It is convenient to discuss large-scale testing problems in terms of the *local false discovery rate* (fdr), an empirical Bayes version of Benjamini and Hochberg's (1995) methodology focusing on densities rather than tail areas (see Efron et al. 2001; Efron and Tibshirani 2002; Storey 2002, 2003).

We begin with a simple Bayes model. Suppose that each of the N z-values falls into one of two classes, "Uninteresting" or "Interesting," corresponding to whether or not z_i is generated according to the null hypothesis, with prior probabilities p_0 and $p_1 = 1 - p_0$ for the classes. Assume that z_i has density either $f_0(z)$ or $f_1(z)$, depending on its class,

$$\begin{aligned} p_0 &= \Pr\{\text{Uninteresting}\}, \quad f_0(z) \text{ density if Uninteresting (Null)}, \\ p_1 &= \Pr\{\text{Interesting}\}, \quad f_1(z) \text{ density if Interesting (Nonnull)}. \end{aligned} \tag{9}$$

The smooth curve in Figure 1 estimates the *mixture density*, $f(z)$,

$$f(z) = p_0 f_0(z) + p_1 f_1(z). \tag{10}$$

According to Bayes theorem, the a posteriori probability of being in the Uninteresting class given z is

$$\Pr\{\text{Uninteresting}|z\} = p_0 f_0(z)/f(z). \tag{11}$$

Here I define the fdr as

$$\mathrm{fdr}(z) \equiv f_0(z)/f(z), \tag{12}$$

ignoring the factor p_0 in (11), so $\mathrm{fdr}(z)$ is an upper bound on $\Pr\{\text{Uninteresting}|z\}$. In fact, p_0 can be roughly estimated (see Remark B in Sec. 7), but I am assuming that p_0 is near 1, say $p_0 \geq .90$, so $\mathrm{fdr}(z)$ is not a flagrant overestimator.

The fdr provides a useful methodology for identifying Interesting cases in a situation like that of Figure 1: (1) estimate $f(z)$ from the observed ensemble of z-values, for example, by the natural spline fit to the histogram counts; (2) assign a null density $f_0(z)$; (3) calculate $\mathrm{fdr}(z) = f_0(z)/f(z)$; and (4) report as Interesting those cases with $\mathrm{fdr}(z_i)$ less than some threshold value, perhaps $\mathrm{fdr}(z_i) \leq .10$. Remark B discusses the close connection between this algorithm and Benjamini and Hochberg's (1995) method.

This article concerns the choice of $f_0(z)$, the null hypothesis density. In the drug mutation example, it is crucial to determine whether f_0 is taken to be the theoretical null, N(0, 1), or the empirical null, $\mathrm{N}(-.35, 1.20^2)$. This is illustrated in Figure 2, a close-up view of Figure 1 focusing on the bin containing $z = -3$. The expected number of the 444 z_i values falling into this bin is 6.37 for $f(z)$, and either .62 or 3.90 as $f_0(z)$

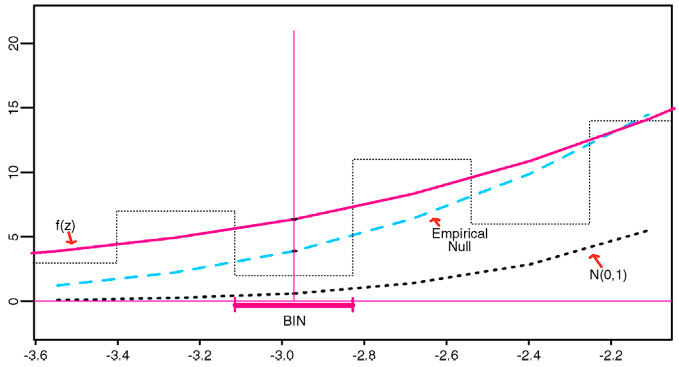

Figure 2. Close-Up View of the Bin Containing $z = -3$ in Figure 1. Expected numbers in the bin are 6.37 for f(z), .62 for $f_0 = N(0, 1)$, and 3.90 for $f_0 = N(.35, 1.20^2)$, the empirical null. Corresponding estimates of fdr(−3) are .097 for N(0, 1) versus .612 for $N(-.35, 1.20^2)$. Should we report the cases in this bin as Interesting?

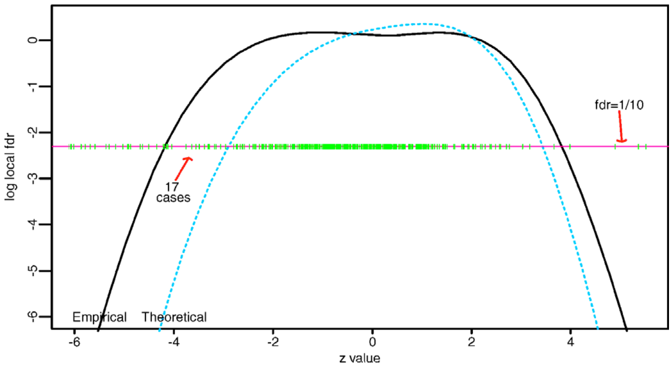

Figure 3. Comparison of Estimates of log fdr(z) for the Drug Mutation Data. The empirical null estimate (——) declines more slowly than the theoretical null estimate (· · · · ·). Dashes indicate the 444 z-values. A total of 17 cases on left have fdr(z) < 1/10 for theoretical but >1/10 for empirical.

is $N(0, 1)$ or $N(-.35, 1.20^2)$. Thus $\text{fdr}(z) = f_0(z)/f(z)$ at $z = -3$ is estimated to be either

$$\text{fdr}(-3) = \begin{cases} .097 & \text{using the theoretical null } N(0, 1) \\ \quad\text{or} & \\ .612 & \text{using the empirical null } N(-.35, 1.20^2). \end{cases} \tag{13}$$

In this bin, changing from the theoretical null to the empirical null changes the inferences from Interesting to definitely Uninteresting.

Figure 3 compares the two estimates of $\log \text{fdr}(z)$ over most of the z scale. As shown, 18 of the 444 z-values have $\text{fdr}(z) < .10$ for $f = N(0, 1)$ but $> .10$ for $f_0 = N(-.35, 1.20^2)$, with 17 of these at the left end of the scale. All told, the empirical null yields only two-thirds as many cases with $\text{fdr} < .10$ as the theoretical null (35 versus 53).

3. ESTIMATING THE EMPIRICAL NULL DISTRIBUTION

The empirical null distribution for the drug mutation data is estimated in two steps: (1) Fit the curve $f(z)$ shown in Figure 1 to the histogram counts by Poisson regression, and (2) Obtain the center and half-width of the central peak, say δ_0 and σ_0, from $f(z)$,

$$\delta_0 = \arg\max\{f(z)\} \quad \text{and} \quad \sigma_0 = \left[-\frac{d^2}{dz^2}\log f(z)\right]_{\delta_0}^{-\frac{1}{2}}, \tag{14}$$

yielding $(\delta_0, \sigma_0) = (-.35, 1.20)$. Details are given in Remark D (Sec. 7), which briefly discusses the possibility of a nonnormal empirical null distribution.

More direct estimation methods for f_0 seem possible; for example, estimating δ_0 by the median of the z-values. Suppose, however, that 10% of the z-values came from the nonnull distribution and that all of these were located at the far left end of Figure 1. Then the median of all the z's would be the 4/9 quartile of the actual null distribution, not its median, yielding a badly biased estimate of δ_0. Similar comments apply to estimating σ_0 (see Remark D). Method (14) does not require preliminary estimates of the proportion p_0 in the null population of (9), a considerable practical advantage.

How accurate are the estimates $(-.35, 1.20)$? The usual standard error approximations for a Poisson regression fit are not appropriate here, because the z_i's are not independent of each other. A nonparametric bootstrap analysis was performed instead, with the 1,391 80-dimensional vectors $(\mathbf{x}, \mathbf{v})$ [(5) and (6)], as the resampling units. This yielded .09 and .08 for the bootstrap standard errors of δ_0 and σ_0, that is,

$$(\delta_0, \sigma_0) = (-.35, 1.20) \pm (.09, .08). \tag{15}$$

It seems quite unlikely that estimation error alone accounts for the difference between the empirical null and the theoretical values $(\delta_0, \sigma_0) = (0, 1)$. (Note that this type of bootstrap analysis, which requires independent sampling units, is not applicable to the microarray example of Sec. 6, where correlations among the genes are present.)

The next two sections concern other possible causes for empirical/theoretical differences, diagnostics for these causes, and their interpretations. This list is not exhaustive, and in fact the microarray example of Section 6 demonstrates another form of pathology.

4. PERMUTATION TESTS AND UNOBSERVED COVARIATES

The theoretical $N(0, 1)$ null hypothesis (4) is usually based on asymptotic approximations like those for the logistic regression coefficients in the drug mutation study. Permutation methods can be used to avoid these approximations, perhaps in the hope that an improved theoretical null will more closely match the empirical.

This was not the case for the drug mutation data, for which permutation testing was implemented by randomly pairing the 1,391 predictor vectors $\mathbf{x}$, (5), with the 1,391 response vectors $\mathbf{v}$, (6), and recalculating the 444 z-values. This whole process was repeated independently 20 times, yielding a total of 20×444 permutation z's. Their distribution was well approximated by a $N(0, .965^2)$ density (the "permutation null"), except for a prominent spike near $z = .3$. In this case, the permutation-improved theoretical null differs more, rather than less, from the empirical null $N(-.35, 1.20^2)$.

Permutation methods are popular in the microarray literature as a way of avoiding assumptions and approximations (see Efron, Tibshirani, Storey, and Tusher 2001; Dudoit, Shaffer, and Boldrick 2003), *but they do not automatically resolve the question of an appropriate null hypothesis.* This can be seen in the following hypothetical example, which is a stylized version of the two-sample microarray testing problem discussed in Section 6. The data, x_{ij}, come from N simultaneous two-sample experiments, each comparing $2n$ subjects,

$$x_{ij}\begin{cases} \text{Controls}, & j = 1, 2, \ldots, n \\ \text{Treatments}, & j = n+1, n+2, \ldots, 2n \end{cases} \qquad (i = 1, \ldots, N). \tag{16}$$

The ith test statistic, Y_i, is the usual two-sample t statistic, comparing Treatments versus Controls for the ith experiment.

Suppose that, unknown to the statistician, the data were actually generated from

$$x_{ij} = u_{ij} + \frac{I_j}{2}\beta_i \qquad \begin{cases} u_{ij} \sim N(0, 1) \\ \beta_i \sim N(0, \sigma_\beta^2), \end{cases} \tag{17}$$

with the u_{ij} and β_i mutually independent and

$$I_j = \begin{cases} -1, & j = 1, 2, \ldots, n \\ +1, & j = n+1, \ldots, 2n \end{cases} \tag{18}$$

Then it is easy to show that the statistics Y_i follow a dilated t distribution with $2n - 2$ df,

$$Y_i \sim \left(1 + \frac{n}{2}\sigma_\beta^2\right)^{\frac{1}{2}} \cdot t_{2n-2}, \tag{19}$$

whereas the permutation distribution, permuting Treatments and Controls within each experiment, has nearly a standard t_{2n-2} null distribution. So, for example, if $\sigma_\beta^2 = 2/n$, then the empirical density of the Y_i's will be $\sqrt{2}$ times as wide as the permutation null.

The quantity β_i in (17) and (18) produces the only consistent differences between Treatments and Controls in experiment i. If β_i is a dependable feature of the ith experiment, and would appear again with the same value in a replication of the study, then the permutation null t_{2n-2} is a reasonable basis for inference. With n large and $\sigma_\beta^2 = 2/n$, this results in $\text{fdr}(y_i) < .10$ for the most extreme 2% of the observed t statistics, favoring those with the largest values of $|\beta_i|$.

Suppose, however, that β_i is not inherent to experiment i, but rather is a purely random effect that would have a different value and perhaps a different sign if the study were repeated; that is, β_i is part of the noise and not part of the signal. In this case, the appropriate choice is the empirical null (19). The equivalent of Figure 1 would be *all* central peak, with no interesting outliers, and no cases with small values of $\text{fdr}(y_i)$. This is appropriate, because now there is no real Treatment effect.

In this latter context β_i acts as an *unobserved covariate*, a quantity that the statistician would use to correct the Treatment–Control comparison if it were observable. Unobserved covariates are ubiquitous in observational studies. There are several obvious ones in the drug mutation study, including personal patient characteristics, such as age and gender, previous use of AZT and other non-PI drugs, years since infection, geographic location, and so on.

The effect of important unobserved covariates is to dilate the null hypothesis density $f_0(z)$, as happens in (19). Unobserved covariates will also dilate the Interesting density $f_1(z)$ in (9), and the mixture density $f(z)$, (10). However, an empirical fitting method for estimating $f(z)$, such as the spline fit in Figure 1, automatically includes any dilation effects. In estimating $\text{fdr}(z) = f_0(z)/f(z)$, it is important to also allow for dilation of the numerator f_0. *This is a strong argument for preferring the empirical null hypothesis in observational studies.*

5. A STRUCTURAL MODEL FOR THE z-VALUES

The Bayesian specifications (9) underlying the fdr results have the advantage of not requiring a structural model for the z-values; in particular, it is not necessary to motivate, or even describe, the nonnull density $f_1(z)$. There is, however, a simple structural model that can help elucidate the Interesting–Uninteresting distinction in (9).

The structural model assumes that z_i, the ith z-value, is normally distributed around a "true value" μ_i, its expectation,

$$z_i \sim \text{N}(\mu_i, 1) \quad \text{for } i = 1, 2, \ldots, N, \tag{20}$$

with μ_i having some prior distribution $g(\mu)$,

$$\mu_i \sim g(\mu) \quad \text{for } i = 1, 2, \ldots, N. \tag{21}$$

Structure (20) is often a good approximation (see Efron 1988, sec. 4), and in fact proved reasonably accurate in the bootstrap experiment yielding (15). Together, (20) and (21) say that the mixture density $f(z)$, (10), is a convolution of $g(\mu)$ with the standard normal density $\varphi(z)$,

$$f(z) = \int_{-\infty}^{\infty} \varphi(z - \mu) g(\mu)\, d\mu \tag{22}$$

[with the understanding that $g(\mu)$ may include discrete probability atoms].

As a first application of the structural model, suppose that we insist that $g(\mu)$ put probability p_0 on $\mu = 0$,

$$\text{Pr}_g\{\mu = 0\} = p_0, \tag{23}$$

for some fixed value of p_0 between 0 and 1. This amounts to the original Bayes model (9) with $p_0 = \text{Pr}\{\text{Uninteresting}\}$, $f_0(z)$ the theoretical null hypothesis N(0, 1), and

$$f_1(z) = \int_{\mu \neq 0} \varphi(z - \mu) g(\mu)\, d\mu \Big/ (1 - p_0). \tag{24}$$

In the context of this article, p_0 should be .90 or greater.

For any $f(z)$ of the convolution form (22), let (δ_g, σ_g) be the center and width parameters (δ_0, σ_0) defined by (14). Figure 4 answers the following question: For a given choice of p_0 in constraint (23), what are the maximum possible values of $|\delta_g|$ and of σ_g,

$$\delta_{\max} = \max\{|\delta_g||p_0\} \quad \text{and} \quad \sigma_{\max} = \max\{\sigma_g|p_0\}? \tag{25}$$

Three curves appear for $\sigma_{\max}$, for the general case just described, for the case where the nonzero component of $g(\mu)$ is required to be symmetric around 0, and for the case where it is also required to be normal. Here only the general case will be discussed. Remark F (Sec. 7) discusses the solution of (25), which turns out to have a simple "single-point" form.

The notable feature of Figure 4 is that for $p_0 \geq .90$, my preferred realm for large-scale hypothesis testing, $(\delta_{\max}, \sigma_{\max})$ must be quite near the theoretical null values (0, 1),

$$\delta_{\max} \leq .07 \quad \text{and} \quad \sigma_{\max} \leq 1.04. \tag{26}$$

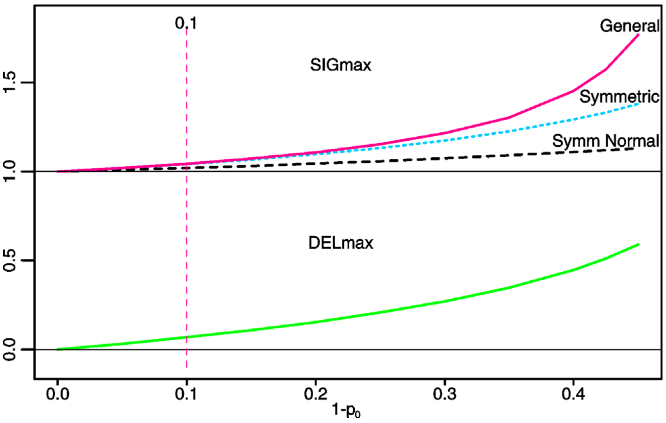

Figure 4. Maximum Possible Values of the Center and Width Parameters (δ_0, σ_0), (14), When the Structural Model (20)–(22) is Constrained to Put Probability p_0 on $\mu = 0$. For $1 - p_0 \leq .10$, the maxima are not much greater than the theoretical null values (0, 1), as shown in Table 1.

Table 1. Value of σ_{max} and δ_{max} as a Function of $1 - p_0$ (23)

$1 - p_0$:	.05	**.10**	.20	.30	(Drug mutation)
σ_{max}:	1.02	**1.04**	1.11	1.22	(1.20)
δ_{max}:	.03	**.07**	.15	.27	(−.35)

Table 1 shows $(\delta_{max}, \sigma_{max})$ for various choices of p_0. It shows that the "Interesting" probability $1 - p_0$ would have to be nearly .30, very large by the standards of large-scale testing, to obtain the observed drug mutation values $(\delta_0, \sigma_0) = (-.35, 1.20)$. The inference is that Uninteresting effects, such as the unobserved covariates of Section 4, are dilating the null hypothesis.

The main point here is that the measures (14) of center and width are quite robust to the arrangement of Interesting values μ_i, as long as the Interesting percentage does not exceed 10%. If (δ_0, σ_0) for the central peak is much different than (0, 1), as it is in Figure 1, then using the theoretical null is bound to result in identifying an uncomfortably large percentage of supposedly Interesting cases.

We can pursue this last point for the drug mutation data by removing constraint (23). Figure 5 shows an unconstrained estimate of $g(\mu)$. For computational simplicity, $g(\mu)$ was assumed to be discrete, with at most $J = 8$ support points $\mu_1, \mu_2, \ldots, \mu_J$, so that (22) becomes

$$f(z) = \sum_{j=1}^{J} \pi_j \varphi(z - \mu_j), \tag{27}$$

π_j being the probability g puts on μ_j, with $\pi_j \geq 0$ and $\sum \pi_j = 1$. A nonlinear minimization program was employed to find the best-fitting curve of form (27) to the histogram counts in Figure 1, using Poisson deviance as the fitting criterion. The vertical bars in Figure 5 are located at the resulting eight values μ_j, with the bar's height proportional to π_j. For example, the little bar at far left represents an atom of probability $\pi_1 = .015$ at $\mu_1 = -10.9$. The resulting $f(z)$ estimate, (26), closely resembles the natural spline fit of Figure 1. Table 2 shows all eight (π_j, μ_j) pairs.

Suppose for a moment that the estimated $g(\mu)$ is exactly correct, so 1.5% of the 444 cases have their μ_i's equal to -10.9, 1.3%, to -7.0, and so on, and that an oracle has told us the eight (π_j, μ_j) values. Given an observed z_i, we can now calculate Pr{Uninteresting$|z$}, (11), exactly, *once the scientist specifies the definition of Uninteresting versus Interesting.* It seems obvious that the 60.8% at $\mu_j = 0$ are Uninteresting, and that the 10.6% at $\mu_j = -10.9$, -7.0, -4.9, and 6.1 deserve Interesting status. However, the status of the 28.6% at $\mu_j = -1.8$, -1.1, and 2.4 is less clear.

If the 28.6% are deemed Interesting, then this leaves only the 60.8% at $\mu_j = 0$ as Uninteresting. In terms of the Bayes model (9), this yields $p_0 = .608$ and $f_0(z) \sim N(0, 1)$, the theoretical null. About 174 of the 444 cases will be identified as Interesting, too many for a typical screening exercise. Shifting the 28.6% to the Uninteresting classification increases p_0 to $.608 + .286 = .894$, a more manageable value, and changes $f_0(z)$ to the version of (27) supported on the four Uninteresting μ_j's,

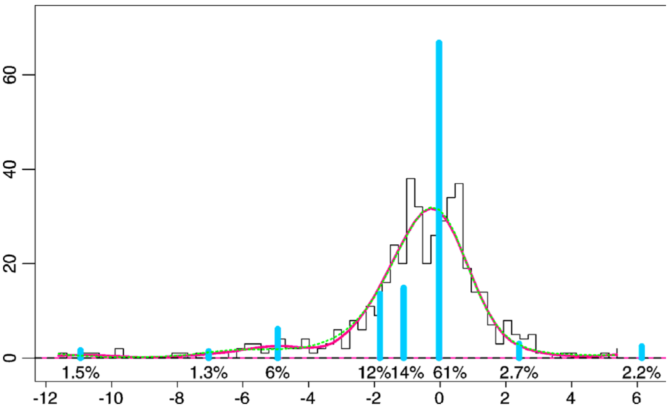

Figure 5. Best-Fit Discrete Mixing Function $g(\mu)$, (21), for Drug Mutation Data. The bars are located at support points μ_j, the heights are proportional to weights π_j, and the tall bar at $\mu_j = 0$ has weight $\pi_j = .61$. Solid curve is a best-fit estimate $f(z) = \sum \pi_j \varphi(z - \mu_j)$; it closely matches the natural spline fit from Figure 1 (- - - -).

$$f_0(z) = \sum_{j=4}^{7} \pi_j \varphi(z - \mu_j) \Big/ \sum_{j=4}^{7} \pi_j. \tag{28}$$

This is approximately $N(-.34, 1.19^2)$, almost the same as the empirical null (8).

In other words, the definition of "Interesting" determines the relevant choice of the null hypothesis f_0. If the goal is to keep the proportion of Interesting cases manageably small, then $f_0(z)$ must grow wider than $N(0, 1)$.

Use of the term "Interesting" rather than "Significant" reflects a difference in intent between large-scale and classical testing. In the hypothetical context of Figure 5 and Table 2, all of the 39.2% of the cases with nonzero μ_i's would eventually be declared as "significantly different from 0" if the sample size of patients was vastly increased. Section 4 suggests that minor deviations from N(0, 1) might arise from scientifically uninteresting causes, such as unobserved covariates. However, even if a modestly nonzero μ_i is genuine in some sense, it may still be Uninteresting when viewed in comparison with an ensemble of more dramatic possibilities. Nonsignificant implies Uninteresting, but not conversely.

6. A MICROARRAY EXAMPLE

Microarrays have become a prime source of large-scale simultaneous testing problems. Figure 6 relates to a well-known

Table 2. Weights π_j and Locations μ_j for the Eight-Point Best-Fit Estimate $g(\mu)$ of Figure 8

		–Interesting–		?	?	Uninteresting	?	Interesting
$100 \cdot \pi_j$	1.5%	1.3%	5.6%	12.3%	13.6%	60.8%	2.7%	2.2%
μ_j	**−10.9**	**−7.0**	**−4.9**	**−1.8**	**−1.1**	**0**	**2.4**	**6.1**

NOTE: Which locations deemed Interesting versus Uninteresting determines the choice between the theoretical or empirical null hypothesis. (Numerical results accurate to one decimal place.)

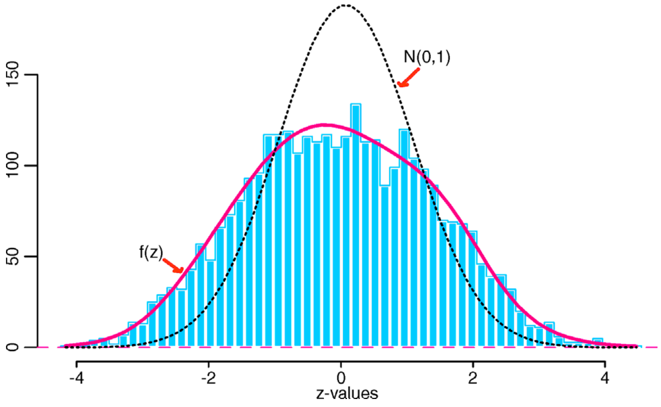

Figure 6. Histogram of $N = 3{,}226$ z-Values From the Breast Cancer Study. The theoretical N(0, 1) null is much narrower than the central peak, which has $(\delta_0, \sigma_0) = (-.02, 1.58)$. In this case the central peak seems to include the entire histogram.

microarray experiment concerning differences between two types of genetic mutations causing increased breast cancer risk, BRCA1 and BRCA2 (see Hedenfalk, Duggen, and Chen 2001; Efron and Tibshirani 2002; Efron 2003).

The experiment included 15 breast cancer patients, 7 with BRCA1 and 8 with BRCA2. Each women's tumor was analyzed on a separate microarray, each microarray reporting on the same set of $N = 3{,}226$ genes. For each gene, the two-sample t statistic y_i comparing the seven BRCA1 responses with the eight BRCA2's was computed. The y_i's were then converted to z-values,

$$z_i = \Phi^{-1} F_{13}(y_i), \tag{29}$$

where F_{13} is the cdf of a standard t distribution with 13 df. Figure 6 displays the histogram of the 3,226 z-values.

The central peak is wider here than in Figure 1, with center-width estimates $(\delta_0, \sigma_0) = (-.02, 1.58)$. More importantly, the histogram seems to be *all* central peak, with no interesting outliers such as those seen at the left of Figure 1. This was reflected in the local fdr calculations; using the theoretical N(0, 1) null yielded 35 genes with $\mathrm{fdr}(z_i) < .1$, those with $|z_i| > 3.35$; using the empirical $\mathrm{N}(-.02, 1.58^2)$ null, no genes at all had fdr $< .1$—or, for that matter, fdr $< .9$, the histogram in fact being a little short-tailed compared with $\mathrm{N}(-.02, 1.58^2)$.

There is ample reason to distrust the theoretical null in this case. The microarray experiment, for all its impressive technology, is still an observational study, with a wide range of unobserved covariates possibly distorting the BRCA1–BRCA2 comparison.

Another reason for doubt can be found in the data itself. The fdr methodology does not require independence of the y_i's or z_i's across genes. However, it does require that the 15 measurements for *each* gene be independent across the microarrays. Otherwise, the two-sample t statistic y_i will not have an t_{13} null distribution, not even approximately.

Unfortunately the experimental methodology used in the breast cancer study seems to have induced substantial correlations among the various microarrays. In particular, as discussed in Remark G, the first four microarrays in the BRCA2 groups were mutually correlated, and likewise the last four. Correlations reduce the effective sample size for a two-sample t statistic, just the type of effect that would induce overdispersion in (29).

This does not say that there are no BRCA1–BRCA2 differences, only that it is dangerous to compare the t statistics with a standard t_{13} null distribution, even if simultaneous inference is accounted for.

7. REMARKS

Remark A (Drug mutation study). The data base for the drug mutation study (Wu et al. 2002), included 2,497 patients having HIV subtype B, of whom 1,391 had received at least one of six popular protease inhibitor (PI) drugs: amprenavir, indinavir, lopinavir, nelfinavir, ritonavir, or saquinavir. Among the 1,391, the mean number of PI drugs taken was 2.05 per patient. Amino acid sequences were obtained at all 99 positions on the HIV protease gene, and mutations from wild-type recorded; 25 positions showed 3 or fewer mutations among the 1,391 patients, deemed too few for analysis, leaving 74 positions for the investigation here. Each of the 74 individual logistic regressions included an intercept term as well as the six PI main effects, but no other covariates.

Remark B (The local false discovery rate). The local fdr, (11) or (12), is closely related to Benjamini and Hochberg's (1995) "tail-area" FDR, as discussed by Efron et al. (2001), Storey (2002), and Efron and Tibshirani (2002). Substituting cdf's F_0 and F for the densities f_0 and f, Bayes's theorem gives a tail-area version of (11),

$$\begin{aligned}\Pr\{\text{Uninteresting}|z \le z_0\} &= p_0 F_0(z_0)/F(z_0)\\ &\equiv \mathrm{FDR}(z_0).\end{aligned} \tag{30}$$

Here $\mathrm{FDR}(z_0)$ turns out to be the conditional expectation of $\mathrm{fdr}(z) \equiv p_0 f_0(z)/f(z)$ given $z \le z_0$,

$$\mathrm{FDR}(z_0) = \int_{-\infty}^{z_0} \mathrm{fdr}(z) f(z)\,dz \bigg/ \int_{-\infty}^{z_0} f(z)\,dz. \tag{31}$$

Benjamini and Hochberg worked in a frequentist framework, but their FDR control rule can be stated in empirical Bayes terms. Given F_0, which they usually took to be what has been called here the theoretical null, they estimate $\mathrm{FDR}(z_0)$ by

$$\widehat{\mathrm{FDR}}(z_0) = p_0 F_0(z)/\widehat{F}(z_0), \tag{32}$$

where $\widehat{F}$ is the empirical cdf of the z_i's. For a desired control level α, say $\alpha = .05$, define

$$z_0 = \arg\max_z \{\widehat{\mathrm{FDR}}(z) \le \alpha\}; \tag{33}$$

then rejecting all cases with $z_i \le z_0$ gives an expected (frequentist) rate of false discoveries no greater than α.

With z_0 as in (33), relation (31) (applied to the estimated versions of FDR, fdr, and f) says that the weighted average of $\mathrm{fdr}(z_i)$ for the cases rejected by the FDR level-α rule is itself α. As an example, take $\alpha = .05$ and f_0 equal the theoretical N(0, 1) null. Applying the FDR control rule to the negative side of Figure 1's drug mutation data rejects the null hypothesis for the 56 cases having $z_i \le -2.61$; the corresponding 56 values of $\mathrm{fdr}(z_i)$ have weighted average $\alpha = .05$. They vary from nearly 0 at the far left to .19 at the boundary value $z = -2.61$, justifying the term "local"; z_i's near the boundary are more likely to be false discoveries than the overall .05 rate suggests.

Our concern with a correct choice of null hypothesis applies to FDR just as well as to fdr. In the microarray study, FDR with $F_0 = \mathrm{N}(0,1)$ gives 24 significant genes at $\alpha = .05$, whereas $F_0 = \mathrm{N}(-.02, 1.58^2)$ gives none. In fact, any simultaneous testing procedure, the popular Westfall–Young method (Westfall and Young, 1993), for example, will depend on a correct assessment of p values for the individual cases, that is, on the choice of F_0.

Remark C [Estimating $f(z)$]. The Poisson regression method used in Figure 1 to estimate the mixture density $f(z)$, (10), originates in an idea of Lindsey described by Efron and Tibshirani (1996, sec. 2). The range of the sample $z_1, z_2, \dots, z_N$ is partitioned into K equal intervals, with interval k having midpoint x_k and containing count s_k of the N z-values; the expectation λ_k of s_k is nearly proportional to $f_k \equiv f(x_k)$, and if the z_i's are independent, then the counts approximate independent Poisson variates,

$$s_k \overset{\text{ind}}{\sim} \mathrm{Poi}(\lambda_k) \quad \text{and} \quad \lambda_k = c f_k, \qquad k = 1, 2, \dots, K, \tag{34}$$

where c is a constant depending on N and the interval length.

Lindsey's method is to estimate the λ_k's with a Poisson regression, which because of (34) amounts to estimating a scaled version of the f_k's; in other words, estimating $f(z)$. K equals 60 in Figure 1, with the regression model being a natural spline with 7 df, roughly equivalent to a sixth degree polynomial fit in z.

Poisson regression based on (34) is almost fully efficient for estimating $f(z)$ if the z_i's are independent. Here one does not expect independence, but we still have the expectation of s_k proportional to f_k. The Poisson regression method will still tend to unbiasedly estimate $f(z)$, assuming the regression model is sufficiently flexible, though we may lose estimating efficiency.

I also used the bootstrap analysis that gave the standard errors in (15) to check (34). This turned out to be surprisingly accurate for the drug mutation data. If it had not, then I might have used the bootstrap estimate of covariance for the s_k's to motivate a more efficient estimation procedure, though this is unlikely to be important for large values of N. In any case bootstrap analyses as in (15) will provide legitimate standard errors for the Poisson regression whether or not (34) is valid.

Remark D (Estimating the empirical null distribution). The main tactic of this article is to estimate the null distribution $f_0(x)$ in (9) from the central peak in the z-values' histogram. Assuming normality for f_0 gives

$$\log f(z) \doteq -\frac{1}{2}\left(\frac{z-\delta_0}{\sigma_0}\right)^2 + \text{constant} \tag{35}$$

for z near 0, so that δ_0 and σ_0 can be estimated by differentiating $\log f(z)$ as in (14). The constant depends on N and p_0, but the constant has no effect on the derivatives in (14).

Directly differentiating the spline estimate of $\log f(z)$ can give an overly variable estimate of σ_0. One more smoothing step was used here, fitting a quadratic curve $a_0 + a_1 x_k + a_2 x_k^2$ by ordinary least squares to the estimated values $\log f_k$, for x_k within 1.5 units of the maximum δ_0, yielding $\sigma_0 = [-2a_2]^{-\frac{1}{2}}$ as in (14). This procedure gave the small bootstrap standard error estimate in (15).

None of this methodology is crucial, although it is important that the estimates δ_0 and σ_0 relate directly to $f_0(z)$ and are not much affected by the nonnull distribution $f_1(z)$ in (9). As an example of what can go wrong, suppose that one tries to estimate σ_0 by a "robust" scale measure, such as (84th quantile minus 16th quantile)/2. This gives $\sigma_0 = 1.47$ for the drug mutation data, reflecting long tails due to the Interesting cases in Figure 1. Similar difficulties arise using the central slope of a qq plot. Basically, a density estimate of the central peak is required, and then some assessment of its center and width.

More ambitiously, one might try extending the estimation of $f_0(z)$ to third moments, permitting a skew null distribution. Expression (35) could be generalized to

$$-\log f(z) \doteq c_0 + c_1 z + c_2 z^2/2 + c_3 z^3/6, \tag{36}$$

now requiring three derivates to estimate the coefficients rather than the two of (14). This is an unexplored path, and in particular Table 1 has not been extended to include skewness bounds.

Familiarity was the only reason for using z-values instead of t-values in Figures 1 and 6.

Remark E (Estimating p_0). One can obtain reasonable upper bounds for p_0 in (9) from estimates of

$$\pi(c) \equiv \Pr_f\{z_i \in \delta_0 \pm c\sigma_0\}. \tag{37}$$

Supposing that $f_0(z) = \mathrm{N}(\delta_0, \sigma_0^2)$, define

$$G_0(c) = 2\Phi(c) - 1 \quad \text{and} \quad G_1(c) = \int_{\delta_0 - c\sigma_0}^{\delta_0 + c\sigma_0} f_1(z)\, dz, \tag{38}$$

the probabilities that $z_i \in \delta_0 \pm c\sigma_0$ under f_0 and f_1. Then

$$p_0 = \frac{\pi(c) - G_1(c)}{G_0(c) - G_1(c)} \le \frac{\pi(c)}{G_0(c)}, \tag{39}$$

the inequality following from the assumption that $G_1(c) \le G_0(c)$; that is, the f_1 density is more dispersed than f_0.

This leads to the estimated upper bound for p_0,

$$\widehat{p}_0 = \frac{\widehat{\pi}(c)}{G_0(c)}, \quad \text{with } \widehat{\pi}(c) = \#\{z_i \in \delta_0 \pm c\sigma_0\}/N. \tag{40}$$

In particular, if it is assumed that $G_1(c) = 0$—in other words, that the Interesting z_i's always fall outside $\delta_0 \pm c\sigma_0$—then $\widehat{p}_0 = \widehat{\pi}(c)/G_0(c)$ is unbiased. (This is the same estimate suggested in remark F of Efron et al. 2001 and Storey 2002.) Choosing $(\delta_0, \sigma_0) = (-.35, 1.20)$ and $c = 1.5$ gave $\widehat{p}_0 = .88$ for the drug mutation data, with bootstrap standard error .024.

Remark F [Single-point solutions for $(\delta_{\max}, \sigma_{\max})$]. The distributions $g(\mu)$ providing $(\delta_{\max}, \sigma_{\max})$ in (25), as graphed in Figure 4, have their nonzero components supported at a single point μ_1. For example, $g(\mu)$ for the entry giving $\sigma_{\max} = 1.04$ in Table 1 puts probability .90 at $\mu = 0$ and .10 at $\mu_1 = 1.47$. Single-point optimality was proved for three of the four cases in Figure 4, and verified by numerical maximization for the "General" case. Here is the proof for the $\sigma_{\max}$ "Symmetric" case; the other two proofs are similar.

Consider symmetric distributions putting probability p_0 on $\mu = 0$ and probabilities p_j on symmetric pairs $(-\mu_j, \mu_j)$, $j = 1, 2, \dots, J$, so (22) becomes

$$f(z) = p_0\varphi(z) + \sum_{j=1}^{J} p_j[\varphi(z-\mu_j) + \varphi(z+\mu_j)]/2. \quad (41)$$

Defining $c_0 = p_0/(1-p_0)$, $r_j = p_j/p_0$, and $r_+ = \sum_1^J r_j = 1/c_0$, σ_0 in (14) can be expressed as

$$\sigma_0 = (1-Q)^{-\frac{1}{2}}, \quad \text{where } Q = \frac{\sum_1^J r_j \mu_j^2 e^{-\mu_j^2/2}}{c_0 r_+ + \sum_1^J r_j e^{-\mu_j^2/2}}. \quad (42)$$

Here $\delta_0 = 0$, which is true by symmetry assuming that $p_0 \geq 1/2$. Then $\sigma_{\max}$ in (25) can be found by maximizing Q.

It will be shown that with p_0 (and c_0) and $\mu_1, \mu_2, \dots, \mu_J$ held fixed in (41), Q is maximized by a choice of $p_1, p_2, \dots, p_J$ having $J-1$ zero values; this is a stronger version of the single-point result. Because Q is homogeneous in $\mathbf{r} = (r_1, r_2, \dots, r_J)$ in (42), the unconstrained maximization of $Q(\mathbf{r})$, subject only to $r_j \geq 0$ for $j = 1, 2, \dots, J$, can be considered.

Differentiation gives

$$\partial Q/\partial r_j = \frac{1}{\text{den}}\left[\mu_j^2 e^{-\mu_j^2/2} - Q \cdot \left(c_0 + e^{-\mu_j^2/2}\right)\right], \quad (43)$$

with "den" the denominator of Q. At a maximizing point $\mathbf{r}$, we must have

$$\frac{\partial Q(\mathbf{r})}{\partial r_j} \leq 0 \quad \text{with equality if } r_j > 0. \quad (44)$$

Defining $R_j = \mu_j^2/(1 + c_0 e^{\mu_j^2/2})$, (43) and (44) yield

$$Q(\mathbf{r}) \geq R_j \quad \text{with equality if } r_j > 0. \quad (45)$$

Because $Q(\mathbf{r})$ is the maximum, this says that r_j, and p_j can be nonzero only if j maximizes R_j. In case of ties, one of the maximizing j's can be arbitrarily chosen.

All of this shows that only $J = 1$ need be considered in (41). The global maximized value of r_0 in (41) is $\sigma_{\max} = (1 - R_{\max})^{-\frac{1}{2}}$, where

$$R_{\max} = \max_{\mu_1}\left\{\mu_1^2/\left(1 + c_0 e^{\mu_1^2/2}\right)\right\}. \quad (46)$$

The maximizing argument μ_1 ranges from 1.43 for $p_0 = .95$ to 1.51 for $p_0 = .70$. The corresponding result for $\delta_{\max}$ is simpler, $\mu_1 = \delta_{\max} + 1$.

Remark G (Microarray correlation in the breast cancer study). It is easy to spot an unwanted correlation structure among the eight BRCA2 microarrays. Let $\mathbf{X}$ be the 3,226 × 8 matrix of BRCA2 data, with the columns of $\mathbf{X}$ standardized to have mean 0 and variance 1. A "de-gened" matrix $\widetilde{\mathbf{X}}$ was formed by subtracting row-wise averages from each element of $\mathbf{X}$,

$$\widetilde{x}_{ij} = x_{ij} - \sum_{k=1}^{8} x_{ik}/8. \quad (47)$$

Table 3 shows the 8 × 8 correlation matrix of $\widetilde{\mathbf{X}}$. With genuine gene effects subtracted out, the correlations should vary around $-1/7 = -.14$ if the columns of $\mathbf{X}$ are independent. Instead, the columns are correlated in blocks of four, with the off-diagonal blocks too negative and the on-diagonal blocks too positive.

Table 3. Correlation Matrix for the BRCA2 Data With Row-Wise Means Subtracted off (46), Indicating Positive Correlations Within the Two Blocks of Four

	1	2	3	4	5	6	7	8
1	1.00	.02	.02	.23	−.36	−.35	−.39	−.34
2	.02	1.00	.10	−.08	−.30	−.30	−.23	−.33
3	.02	.10	1.00	−.17	−.21	−.26	−.31	−.27
4	.23	−.08	−.17	1.00	−.30	−.23	−.27	−.32
5	−.36	−.30	−.21	−.30	1.00	−.02	.11	.22
6	−.35	−.30	−.26	−.23	−.02	1.00	.15	.13
7	−.39	−.23	−.31	−.27	.11	.15	1.00	.07
8	−.34	−.33	−.27	−.32	.22	.13	.07	1.00

Remark H (Scaling properties). The associate editor pointed out that the combination of empirical null hypotheses with false discovery rate methodology "scales up" nicely, in terms of both the number of tests and the amount of information per test. Consider the structural model (20), (21) with $g(\mu)$ a mixture of 99% $\mu \sim \mathrm{N}(0, .01)$ and 1% of $\mu = 5$. For N the number of tests large enough, methods like Bonferroni bounds that control the family-wise error rate will eventually accept all N null hypotheses; fdr methods, using either the empirical or theoretical null, will correctly identify most of the Interesting 1%.

Suppose now that the amount of information per test increases by a factor of n, so that each $\mu_i \to \sqrt{n}\,\mu_i$ in (21). Using the theoretical N(0, 1) null makes fdr reject all N cases for n sufficiently large, whereas the empirical null continues to identify only the Interesting 1%. In this context, the fdr/empirical combination avoids the standard criticism of hypotheses testing, that rejection becomes certain for large sample sizes.

8. SUMMARY

Large-scale simultaneous hypothesis testing, where the number of cases exceeds, say 100, permits the empirical estimation of a null hypothesis distribution. The empirical null may be wider (more dispersed) than the theoretical null distribution that would ordinarily be used for a single hypothesis test. The choice between empirical and theoretical nulls can greatly influence which cases are identified as "Significant" or "Interesting," as opposed to "Null" or "Uninteresting," this being true no matter which simultaneous hypothesis testing method is used.

This article presents an analysis plan for large-scale testing situations:

- A density fitting technique is used to estimate the null hypothesis distribution f_0, (Fig. 1 and Sec. 3).
- The local false discovery rate (fdr), an empirical Bayes version of standard FDR theory, provides inferences for the N cases (Fig. 3 and Sec. 2).

There are many possible reasons for overdispersion of the empirical null distribution that would lead to the empirical null being preferred for simultaneous testing including:

- Unobserved covariates in an observational study, (Sec. 4)
- Hidden correlations (Sec. 6)
- A large proportion of genuine but uninterestingly small effects (Fig. 5).

Large-scale testing differs in scientific intent from an individual hypothesis test. The latter is most often designed to reject the null hypothesis with high probability. Large-scale testing is usually more of a screening operation, intended to identify a *small* percentage of Interesting cases, assumed to be on the order of 10% or less in this article. The empirical null hypothesis methodology is designed to be accurate under this constraint (Fig. 4). More traditional estimation methods, involving permutations or quantiles, give incorrect f_0 estimates (Sec. 4 and Remark D).

[Received June 2003. Revised August 2003.]

REFERENCES

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.

Dudoit, S., Shaffer J., and Boldrick J. (2003), "Multiple Hypothesis Testing in Microarray Experiments," *Statistical Science*, 18, 71–103.

Efron, B. (1988), "Three Examples of Computer-Intensive Statistical Inference," *Sankhyā*, 50, 338–362.

——— (2003), "Robbins, Empirical Bayes, and Microarrays," *The Annals of Statistics*, 31, 366–378.

Efron, B., and Tibshirani, R. (1996), "Using Specially Designed Exponential Families for Density Estimation," *The Annals of Statistics*, 24, 2431–2461.

——— (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," *Journal of the American Statistical Association*, 96, 1151–1160.

Hedenfalk, I., Duggen, D., Chen, Y. et al. (2001), "Gene Expression Profiles in Hereditary Breast Cancer," *New England Journal of Medicine*, 344, 539–548.

Miller, R. (1981), *Simultaneous Statistical Inference* (2nd ed.), New York: Springer-Verlag.

Storey, J. (2002), "A Direct Approach to False Discovery Rates," *Journal of the Royal Statistical Society*, Ser. B, 64, 479–498.

——— (2003), "The Positive False Discovery Rate: A Bayesian Interpretation and the q-Value," *The Annals of Statistics*, 31, to appear.

Westfall, P., and Young, S. (1993), *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustments*, New York: Wiley.

Wu, T., Schiffer, C., Shafer, R. et al. (2003), "Mutation Patterns and Structural Correlates in Human Immunodeficiency Virus Type 1 Protease Following Different Protease Inhibitor Treatments," *Journal of Virology*, 77, 4836–4847.