Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: http://www.tandfonline.com/loi/uasa20

Correlated z-Values and the Accuracy of Large-Scale Statistical Estimates

Bradley Efron

To cite this article: Bradley Efron (2010) Correlated z-Values and the Accuracy of Large-Scale Statistical Estimates, Journal of the American Statistical Association, 105:491, 1042-1055, DOI: 10.1198/jasa.2010.tm09129

To link to this article: http://dx.doi.org/10.1198/jasa.2010.tm09129

Correlated *z*-Values and the Accuracy of Large-Scale Statistical Estimates

Bradley EFRON

We consider large-scale studies in which there are hundreds or thousands of correlated cases to investigate, each represented by its own normal variate, typically a z-value. A familiar example is provided by a microarray experiment comparing healthy with sick subjects' expression levels for thousands of genes. This paper concerns the accuracy of summary statistics for the collection of normal variates, such as their empirical cdf or a false discovery rate statistic. It seems like we must estimate an N by N correlation matrix, N the number of cases, but our main result shows that this is not necessary: good accuracy approximations can be based on the root mean square correlation over all $N \cdot (N-1)/2$ pairs, a quantity often easily estimated. A second result shows that z-values closely follow normal distributions even under nonnull conditions, supporting application of the main theorem. Practical application of the theory is illustrated for a large leukemia microarray study.

KEY WORDS: Acceleration; Correlation penalty; Empirical process; Mehler's identity; Nonnull z-values; Rms correlation.

1. INTRODUCTION

Modern scientific studies routinely produce data on thousands of related situations. A familiar example is a microarray experiment in which thousands of genes are being investigated for possible disease involvement. Each gene might produce a z-value, say z_i, for the ith gene, by definition a test statistic theoretically having a standard normal distribution

$$H_0 : z_i \sim \mathcal{N}(0, 1) \tag{1.1}$$

under the null hypothesis H_0 of no disease involvement. A great deal of the current literature was developed under the assumption of independence among the z_i's. This can be grossly unrealistic in practice, as discussed in Owen (2005) and Efron (2007a), among others. This paper concerns the accuracy of summary statistics of the z_i's, for example, their empirical cumulative distribution function (cdf), under conditions of substantial correlation.

Figure 1 concerns a leukemia microarray study by Golub et al. (1999) that we will use for motivation and illustration. Two forms of leukemia are being examined for possible genetic differences: acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML). In the version of the data discussed here there are $n_1 = 47$ ALL patients and $n_2 = 25$ AML patients, with expression levels on the same $N = 7128$ genes measured on each patient.

A two-sample t-statistic t_i comparing AML with ALL expression levels was computed for each gene and converted to a z-value,

$$z_i = \Phi^{-1}(F_{70}(t_i)), \qquad i = 1, 2, \ldots, N, \tag{1.2}$$

where Φ and F_{70} are the cumulative distribution functions for a standard normal and a Student-t distribution with 70 degrees of freedom. Figure 1 shows a histogram of the z_i's, which turns out to be much wider than (1.1) suggests: its central spread is estimated to be $\hat{\sigma}_0 = 1.68$ rather than 1, as discussed in Section 3.

Here is an example of the results to be derived in Sections 2 through 4. Let $\hat{F}(x)$ be the right-sided cdf ("survival curve") of the z-values,

$$\hat{F}(x) = \#\{z_i > x\}/N. \tag{1.3}$$

Then a good approximation for the variance of $\hat{F}(x)$ is

$$\operatorname{Var}\{\hat{F}(x)\} \doteq \left\{\frac{\hat{F}(x)(1-\hat{F}(x))}{N}\right\} + \left\{\frac{\hat{\sigma}_0^2 \hat{\alpha} \hat{f}^{(1)}(x)}{\sqrt{2}}\right\}^2. \tag{1.4}$$

The first term in (1.4) is the usual binomial variance, while the second term is a *correlation penalty* accounting for dependence between the z_i's. The quantities occurring in the correlation penalty are

- $\hat{\sigma}_0$, the estimate of central spread (1.68 above);
- $\hat{\alpha}$, an estimate of the root-mean-square of the correlations between the $N(N-1)/2$ pairs of z_i's (equaling about 0.11 for the leukemia data, as calculated from the simple formula in Section 3);
- $\hat{f}^{(1)}(x)$, the first derivative of a smooth fit to the z-value histogram (estimated by a Poisson spline regression in Figure 1).

The row marked $\widehat{\text{sd}}$ in Table 1 is the square root of formula (1.4) applied to the leukemia data. $\hat{F}(4) = 0.025$ is seen to have $\widehat{\text{sd}} = 0.0040$, more than double $\widehat{\text{sd}}_0 = 0.0018$, the binomial standard deviation obtained by ignoring the second term in (1.4). The permutation standard deviation, obtained from repeated permutations of the 72 patients, is only 0.0001 at $x = 4$. Permutation methods, which preserve within-microarray correlations, have been advocated for large-scale hypothesis testing (see Westfall and Young 1993, Dudoit, Shaffer, and Boldrick 2003, section 2.6), but they are inappropriate for the accuracy considerations of this paper.

Formula (1.4), and more ambitious versions that include covariances across different values of x, are derived in Sections 2 and 3: an exact expression is derived first, followed by a series of simplifying approximations and techniques for their estimation. The basic results are extended to provide more general

Bradley Efron is Professor, Department of Statistics, Stanford University, Stanford, CA 94305-4065 (E-mail: *brad@stat.stanford.edu*). This work was supported in part by NIH grant 8R01 EB002784 and NSF grant DMS0505673.

Journal of the American Statistical Association
September 2010, Vol. 105, No. 491, Theory and Methods
DOI: 10.1198/jasa.2010.tm09129

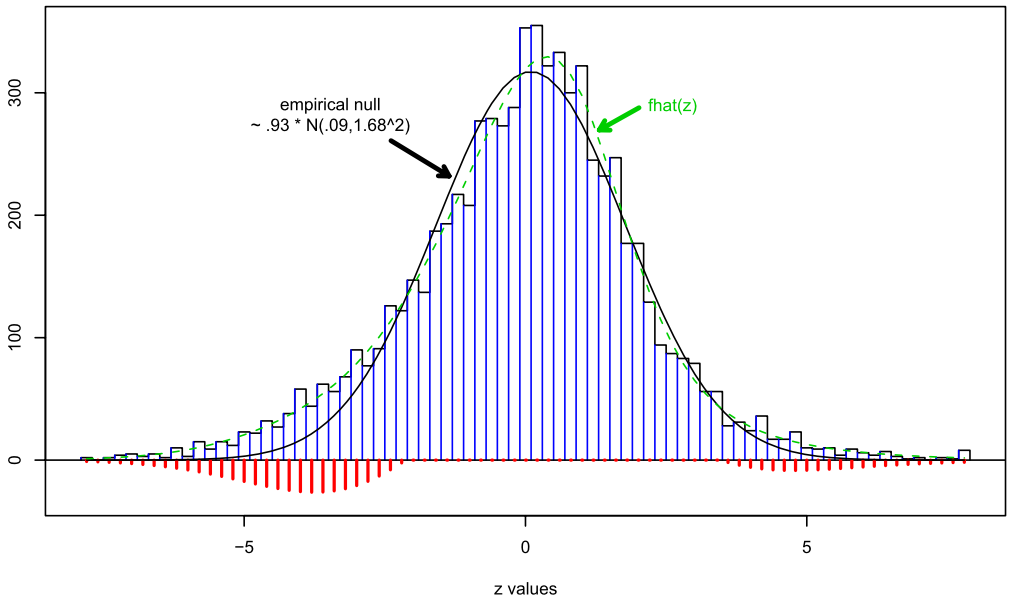

Figure 1. Histogram of z-values for $N = 7128$ genes, leukemia study, Golub et al. (1999). *Dashed curve* $\hat{f}(x)$, smooth fit to histogram; *solid curve* "empirical null," normal density fit from central 50% of histogram, is much wider than theoretical $\mathcal{N}(0, 1)$ null distribution. Small red bars plotted negatively discussed in Section 4. The online version of this figure is in color.

accuracy estimates in Section 4: comparing, for example, the variability of local versus tail-area false discovery rates.

All of our results depend on the assumption that the z_i's are normal with possibly different means and variances,

$$z_i \sim \mathcal{N}(\mu_i, \sigma_i^2), \qquad i = 1, 2, \ldots, N. \tag{1.5}$$

There is no requirement that they be "z-values" in the hypothesis-testing sense of (1.1), and in fact this paper is more concerned with estimation than testing. However, z-values are ubiquitous in large-scale applications, and not only in the two-sample setting of the leukemia study. Section 5 concerns the *nonnull distribution of z-values*. A theorem is derived justifying (1.5) as a good approximation, allowing results like (1.4) to be applied to the leukemia study z-values. Sections 6 and 7 close with remarks and a brief summary.

The statistics microarray literature has shown considerable interest in the effects of large-scale correlation, some good references being Dudoit, van der Laan, and Pollard (2004), Owen (2005), Qiu, Klebanov, and Yakovlev (2005), Qiu et al. (2005), and Desai, Deller, and McCormick (2009). Efron (2007a) used a z-value setting to examine the effects of correlation on false discovery rate analysis; that paper's Section 2 theorem is a null hypothesis version of the general result developed here. A useful extension along different lines appears in Schwartzman and Lin (2009).

Clarke and Hall's (2009) asymptotic calculations support the use of the independence standard deviation $\widehat{\text{sd}}_0$ in Table 1, even in the face of correlation. The situations they consider are low-correlation by the standard here, with the root-mean-square value $\hat{\alpha}$ of (1.4) approaching zero [from their assumption (3.2)]. Since $\hat{\alpha}$ is often easy to estimate, formulas such as (1.4) provide a quantitative check on the use of $\widehat{\text{sd}}_0$.

Table 1. Estimates of standard deviation for right-sided cdf $\hat{F}(x)$ (1.3); $\widehat{\text{sd}}$ square root of formula (1.4); $\widehat{\text{sd}}_0$ square root of first term in (1.4); $\widehat{\text{sd}}_{\text{perm}}$ permutation standard deviation. Accuracy of False Discovery Rate estimate $\widehat{\text{Fdr}}(x)$ discussed in Section 4

	x				
	1	2	3	4	5
$\widehat{\text{sd}}$	**0.017**	**0.022**	**0.0101**	**0.0040**	**0.0019**
$\widehat{\text{sd}}_0$	0.005	0.004	0.0027	0.0018	0.0012
$\widehat{\text{sd}}_{\text{perm}}$	0.021	0.001	0.0014	0.0001	0.0000
$\hat{F}(x)$	0.29	0.13	0.057	0.025	0.010
$\widehat{\text{Fdr}}(x)$	0.94	0.92	0.71	0.38	0.15

2. THE DISTRIBUTION OF CORRELATED NORMAL VARIATES

Given N correlated normal variates $z_1, z_2, \ldots, z_N$, with possibly different means and standard deviations, let $\hat{F}(x)$ denote their right-sided empirical cdf (it is convenient for the applications of Section 4 to deal with right-sided cdfs or *survival curves* instead of the usual left-sided ones in (1.2), and we will use this definition in what follows)

$$\hat{F}(x) = \#\{z_i \geq x\}/N \quad \text{for } -\infty < x < \infty. \tag{2.1}$$

This section presents tractable formulas for the mean and covariance of the process $\{\hat{F}(x),\ -\infty < x < \infty\}$, and a simpler approximation that we will see is nicely suited for applications.

Rather than work directly with cdfs, it will be easier and, in a sense, more basic to first derive results for a discretized version of the empirical *density* of the z_i values. We partition the range $\mathcal{Z}$ into K bins $\mathcal{Z}_k$,

$$\mathcal{Z} = \bigcup_{k=1}^{K} \mathcal{Z}_k, \tag{2.2}$$

each bin being of width Δ. Let x_k indicate the midpoint of $\mathcal{Z}_k$, and y_k the number of z_i's in $\mathcal{Z}_k$,

$$y_k = \#\{z_i \in \mathcal{Z}_k\}, \qquad k = 1, 2, \ldots, K. \tag{2.3}$$

We will derive expressions for the mean and covariance of the vector $\mathbf{y} = (y_1, y_2, \ldots, y_K)'$. In effect, $\mathbf{y}$ is the order statistic of $\mathbf{z} = (z_1, z_2, \ldots, z_N)'$, becoming exactly that as the bin width $\Delta \to 0$. (In which case the y_k values go to 1 or 0, with the nonzero bin x_k values indicating the locations of the ordered z_i's, assuming no ties.) Familiar statistical applications, of the type described in Section 4, depend on $\mathbf{z}$ only through $\mathbf{y}$.

Suppose that the z_i's are divided into a finite number of classes, with members of the cth class $\mathcal{C}_c$ having mean μ_c and standard deviation σ_c,

$$z_i \sim \mathcal{N}(\mu_c, \sigma_c^2) \quad \text{for } z_i \in \mathcal{C}_c. \tag{2.4}$$

Let N_c be the number of members of $\mathcal{C}_c$, with p_c the proportion

$$N_c = \#\{\mathcal{C}_c\} \qquad \text{and} \qquad p_c = N_c/N \tag{2.5}$$

so $\sum_c N_c = N$ and $\sum_c p_c = 1$. The use of model (2.4) for z-values is supported by the results of Section 5.

If $\mathbf{x}$ is the K-vector of bin midpoints, let $x_{kc} = (x_k - \mu_c)/\sigma_c$ and

$$\mathbf{x}_c = (\mathbf{x} - \mu_c)/\sigma_c = (\ldots, x_{kc}, \ldots)'. \tag{2.6}$$

Likewise, for any real-valued function $h(x)$ we define $\mathbf{h}_c$ to be the K-vector of function values

$$\mathbf{h}_c = (\ldots, h(x_{kc}), \ldots)', \tag{2.7}$$

also denoted by $h(\mathbf{x}_c)$ in what follows.

It is easy to calculate the expectation of the count vector $\mathbf{y}$ under the multiclass model (2.4)–(2.5). Let π_{kc} equal the probability that z_i from class $\mathcal{C}_c$ falls into the kth bin,

$$\pi_{kc} = \text{Prob}_c\{z_i \in \mathcal{Z}_k\} \doteq \Delta\varphi(x_{kc})/\sigma_c. \tag{2.8}$$

Here $\varphi(x) = \exp(-x^2/2)/\sqrt{2\pi}$, the standard normal density. The approximation $\pi_{kc} \doteq \Delta\varphi(x_{kc})/\sigma_c$ from (2.4) becomes arbitrarily accurate for Δ sufficiently small, and we will take it as exact in what follows. Then

$$\begin{aligned} E\{\mathbf{y}\} &= N\sum_c p_c \boldsymbol{\pi}_c = N\Delta \sum_c p_c \varphi(\mathbf{x}_c)/\sigma_c \\ &= N\Delta \sum_c p_c \boldsymbol{\varphi}_c/\sigma_c. \end{aligned} \tag{2.9}$$

The $K \times K$ covariance matrix of the count vector $\mathbf{y}$ depends on the $N \times N$ correlation matrix of $\mathbf{z}$, but in a reasonably simple way discussed next. Two important definitions are needed to state the first result: there are $M = N(N-1)/2$ correlations $\rho_{ii'}$ between pairs $(z_i, z_{i'})$ of members of $\mathbf{z}$, and we denote by "$g(\rho)$" the distribution putting weight $1/M$ on each $\rho_{ii'}$. Also, for $\varphi_\rho(u, v)$ the bivariate normal density having zero means, unit standard deviations, and correlation ρ, we define

$$\begin{aligned} &\lambda_\rho(u, v) \\ &\quad = \frac{\varphi_\rho(u, v)}{\varphi(u)\varphi(v)} - 1 \\ &\quad = (1-\rho^2)^{-1/2} \exp\left\{\frac{2\rho uv - \rho^2(u^2+v^2)}{2(1-\rho^2)}\right\} - 1 \end{aligned} \tag{2.10}$$

and

$$\lambda(u, v) = \int_{-1}^{1} \lambda_\rho(u, v) g(\rho)\, d\rho \tag{2.11}$$

(the integral notation being shorthand for summing over M discrete points).

Lemma 1. Under the multiclass model (2.4)–(2.5), the covariance of the count vector $\mathbf{y}$ (2.3) has two components,

$$\mathbf{cov}(\mathbf{y}) = \mathbf{cov}_0 + \mathbf{cov}_1, \tag{2.12}$$

where

$$\mathbf{cov}_0 = N\sum_c p_c\{\text{diag}(\boldsymbol{\pi}_c) - \boldsymbol{\pi}_c\boldsymbol{\pi}_c'\} \tag{2.13}$$

and

$$\begin{aligned} \mathbf{cov}_1 = N^2 \sum_c \sum_d p_c p_d \,\text{diag}(\boldsymbol{\pi}_c)\boldsymbol{\lambda}_{cd}\,\text{diag}(\boldsymbol{\pi}_d) \\ - N\sum_c p_c \,\text{diag}(\boldsymbol{\pi}_c)\boldsymbol{\lambda}_{cc}\,\text{diag}(\boldsymbol{\pi}_c). \end{aligned} \tag{2.14}$$

Here $\text{diag}(\boldsymbol{\pi}_c)$ is the $K \times K$ diagonal matrix having diagonal elements π_{kc}, similarly $\text{diag}(\boldsymbol{\pi}_d)$, while $\boldsymbol{\lambda}_{cd}$ is the $K \times K$ matrix with klth element $\lambda(x_{kc}, x_{ld})$; the summations are over all classes.

Note 1. Equation (2.14) assumes that the correlation distribution $g(\rho)$ is the same across all classes $\mathcal{C}_c$. The proof of Lemma 1, which is similar to that for the simpler situation of Efron (2007a), appears in Remark C of Section 6.

The $\mathbf{cov}_0$ term in (2.12)–(2.13) is the sum of the multinomial covariance matrices that would apply if the z_i's were mutually independent with fixed numbers drawn from each class; $\mathbf{cov}_1$ is a penalty for correlation, almost always increasing $\mathbf{cov}(\mathbf{y})$. The N^2 factor in (2.14) makes the correlation penalty more severe as N increases, assuming $g(\rho)$ stays the same.

Expression (2.14) for the correlation penalty can be considerably simplified. *Mehler's identity* for $\lambda_\rho(u, v)$ (2.10) is

$$\lambda_\rho(u, v) = \sum_{j\ge1} \frac{\rho^j}{j!} h_j(u) h_j(v), \tag{2.15}$$

where h_j is the jth Hermite polynomial. [See Lancaster 1958 for an enlightening discussion of (2.15), also known as the "tetrachoric series," and its connections to the singular value decomposition, canonical correlation, Pearson's coefficient of contingency, and correspondence analysis.] Denoting the jth moment of the correlation distribution $g(\rho)$ by α_j,

$$\alpha_j = \int_{-1}^{1} \rho^j g(\rho)\, d\rho, \tag{2.16}$$

(2.11) becomes

$$\lambda(u, v) = \sum_{j\ge1} \frac{\alpha_j}{j!} h_j(u) h_j(v) \tag{2.17}$$

so $\boldsymbol{\lambda}_{cd}$ in (2.14) can be written in outer product notation as

$$\boldsymbol{\lambda}_{cd} = \sum_{j\ge1} \frac{\alpha_j}{j!} h_j(\mathbf{x}_c) h_j(\mathbf{x}_d)'. \tag{2.18}$$

Making use of (2.8), taken as exact,

$$\begin{aligned}\operatorname{diag}(\boldsymbol{\pi}_c)h_j(\mathbf{x}_c) &= N\Delta \operatorname{diag}(\varphi(\mathbf{x}_c))h_j(\mathbf{x}_c)/\sigma_c \\ &= (-1)^j N\Delta \cdot \boldsymbol{\varphi}_c^{(j)}/\sigma_c, \end{aligned} \tag{2.19}$$

where $\boldsymbol{\varphi}_c^{(j)}$ indicates the jth derivative of $\varphi(u)$ evaluated at each component of $\mathbf{x}_c$ [using $\varphi^{(j)}(u) = (-1)^j\varphi(u)h_j(u)$].

Rearranging (2.14) then gives a simplified formula.

Lemma 2. Defining

$$\bar{\boldsymbol{\phi}}^{(j)} \equiv \sum_c p_c \boldsymbol{\varphi}_c^{(j)}/\sigma_c, \tag{2.20}$$

(2.14) for the correlation penalty becomes

$$\mathbf{cov}_1 = N^2\Delta^2\left\{\sum_{j\ge1}\frac{\alpha_j}{j!}\bar{\boldsymbol{\phi}}^{(j)}\bar{\boldsymbol{\phi}}^{(j)\prime} - \frac{1}{N}\sum_{j\ge1}\frac{\alpha_j}{j!}\left(\sum_c p_c\boldsymbol{\varphi}_c^{(j)}\boldsymbol{\varphi}_c^{(j)\prime}/\sigma_c^2\right)\right\}. \tag{2.21}$$

A convenient approximation to $\mathbf{cov}_1$ is based on three reductions of (2.21):

- The second term in (2.21) is neglible for large N.
- Common standardization methods for large-scale datasets often make α_1, the expectation of $g(\rho)$, exactly or nearly zero, as illustrated in Section 3 for the leukemia data; see section 3 of Efron (2009).
- This leaves α_2 of (2.16) as the leading term in (2.21). With ρ confined to $[-1, 1]$, the higher-order moments $\alpha_j = E_g\{\rho^j\}$ often decrease quickly to zero.

The root mean square (rms) correlation

$$\alpha = \alpha_2^{1/2} = \left[\int_{-1}^{1}\rho^2 g(\rho)\,d\rho\right]^{1/2} \tag{2.22}$$

featured in Efron (2007a) (where it is called the *total correlation*), is a single-number summary of $\mathbf{z}_i$'s entire correlation structure. Carrying out the three reductions above produces a greatly simplified form of (2.21),

$$\textit{rms approximation}: \quad \mathbf{cov}_1 \doteq (N\Delta\alpha)^2\bar{\boldsymbol{\phi}}^{(2)}\bar{\boldsymbol{\phi}}^{(2)\prime}/2 \tag{2.23}$$

with $\bar{\boldsymbol{\phi}}^{(2)}$ in (2.20) depending on the second derivative of the normal density, $\varphi^{(2)}(u) = \varphi(u)\cdot(u^2-1)$.

Figure 2 compares the exact formulas (2.12)–(2.14) for $\mathbf{cov}(\mathbf{y})$ with the simplified formula based on the rms approximation (2.23); for a numerical example having $N = 6000$, $\alpha = 0.10$, and two classes (2.4)–(2.5), initially with

$$\begin{aligned}(\mu_0, \sigma_0) &= (0, 1), \qquad p_0 = 0.95 \qquad \text{and} \\ (\mu_1, \sigma_1) &= (2.5, 1), \qquad p_1 = 0.05\end{aligned} \tag{2.24}$$

but then recentered as in the leukemia example; see Remark D of Section 6 for more detail. The plotted curves show the standard deviations $\mathrm{sd}\{y_k\} = \mathrm{cov}_{kk}(\mathbf{y})^{1/2}$ from (2.12), the corresponding rms approximation (2.23), and also $\mathrm{sd}_0\{y_k\} = (\mathrm{cov}_{0kk})^{1/2}$ from (2.13). We can see there is a substantial correlation penalty over most of the range of z, and also that the rms approximation is quite satisfactory here.

Returning to right-sided cdfs (2.1), let $\mathbf{B}$ be the $K\times K$ matrix

$$\mathbf{B}_{kk'} = \begin{cases} 1 & \text{if } k \le k' \\ 0 & \text{if } k > k', \end{cases} \tag{2.25}$$

so

$$\hat{\mathbf{F}} = \frac{1}{N}\mathbf{B}\mathbf{y} \tag{2.26}$$

is a K-vector with kth component the proportion of z_i's in bins indexed $\ge k$,

$$\hat{F}_k = \#\{z_i \ge x_k - \Delta/2\}/N \qquad (k = 1, 2, \ldots, K). \tag{2.27}$$

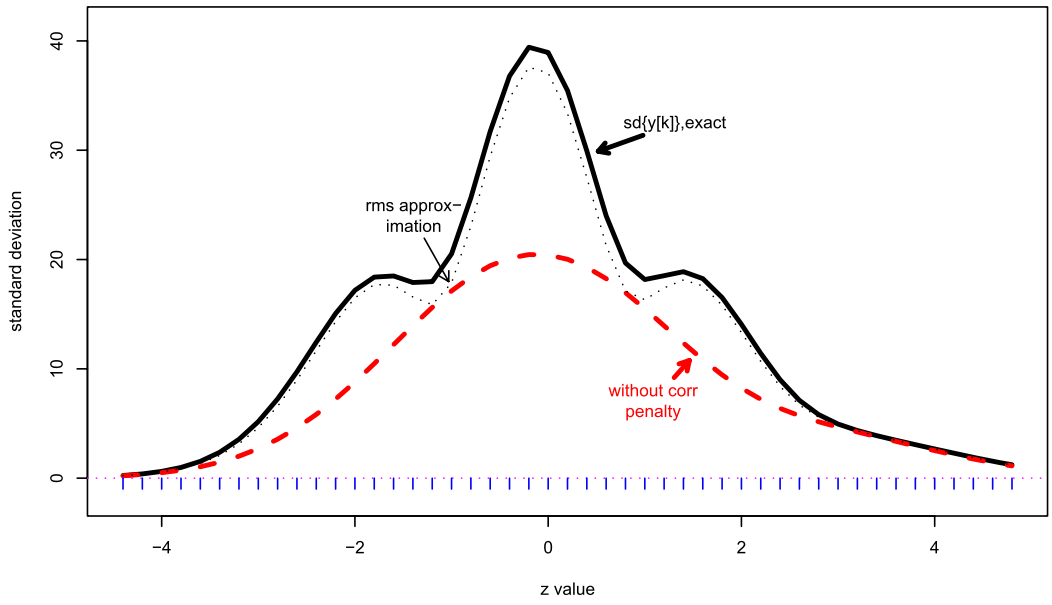

Figure 2. Comparison of exact formula for standard deviation of y_k from (2.12) (heavy curve) with rms approxmation from (2.23) (dotted curve); $N = 6000$, $\alpha = 0.10$ in (2.22), two classes as in (2.24). Dashed curve is standard deviation from (2.13) ignoring the correlation penalty. Hash marks indicate bin midpoints x_k. The online version of this figure is in color.

(**B** would be transposed if we were dealing with left-sided cdfs.) The expectation of $\hat{\mathbf{F}}$ is both obvious and easy to obtain from (2.9),

$$\begin{aligned} E\{\hat{F}_k\} &= \sum_c p_c \left[\sum_{k' \geq k} \Delta\varphi\left(\frac{x_{k'} - \mu_c}{\sigma_c}\right)/\sigma_c \right] \\ &\doteq \sum_c p_c \int_{x_{kc}}^{\infty} \varphi(u)\, du \\ &= \sum_c p_c \Phi^+(x_{kc}), \end{aligned} \tag{2.28}$$

where $\Phi^+(u) = 1 - \Phi(u)$. Now that we are working with tail areas rather than densities we can let $\Delta \to 0$, making (2.28) exact.

$\hat{\mathbf{F}}$ has covariance matrix $\mathbf{Bcov}(\mathbf{y})\mathbf{B}'/N^2$. The same kind of calculations as in (2.28) applied to Lemma 1 gives the following theorem.

Theorem 1. Under the multiclass model (2.4)–(2.5),

$$\mathbf{Cov}(\hat{\mathbf{F}}) = \mathbf{Cov}_0 + \mathbf{Cov}_1, \tag{2.29}$$

where $\mathbf{Cov}_0$ has klth entry

$$\frac{1}{N} \sum_c p_c \left\{ \Phi^+(\max(x_{kc}, x_{lc})) - \Phi^+(x_{kc})\Phi^+(x_{lc}) \right\} \tag{2.30}$$

and

$$\mathbf{Cov}_1 = \sum_j \frac{\alpha_j}{j!} \bar{\boldsymbol{\varphi}}^{(j-1)} \bar{\boldsymbol{\varphi}}^{(j-1)\prime} - \frac{1}{N} \sum_j \frac{\alpha_j}{j!} \left\{ \sum_c p_c \boldsymbol{\varphi}_c^{(j-1)} \boldsymbol{\varphi}_c^{(j-1)\prime} \right\}. \tag{2.31}$$

Here p_c is from (2.5), x_{kc} and x_{lc} from (2.6), α_j is as in (2.16) and

$$\bar{\boldsymbol{\varphi}}^{(j-1)} = \sum_c p_c \boldsymbol{\varphi}_c^{(j-1)} = \sum_c p_c \varphi^{(j-1)}(\mathbf{x}_c). \tag{2.32}$$

[Notice the distinction between $\bar{\boldsymbol{\varphi}}$ and $\bar{\boldsymbol{\phi}}$ (2.20), and between **Cov** and **cov** etc., Lemma 1.]

The three-step reduction leading to (2.31) also can be applied to $\mathbf{Cov}_1$: for α as in (2.22),

$$\textit{Rms approximation}: \quad \mathbf{Cov}_1 \doteq \alpha^2 \bar{\boldsymbol{\varphi}}^{(1)} \bar{\boldsymbol{\varphi}}^{(1)\prime}/2 \tag{2.33}$$

with $\bar{\boldsymbol{\varphi}}^{(1)}$ depending on the first derivative of the normal density, $\varphi^{(1)}(u) = -\varphi(u)u$. Section 3 shows that (2.33) is especially convenient for applications.

Figure 3 is the version of Figure 2 applying to $\hat{\mathbf{F}}$: the heavy curve tracks $\mathrm{sd}(\hat{F}_k)$ from (2.29), the dotted curve is from Rms approximation (2.33), and the dashed curve shows the standard deviations from $\mathbf{Cov}_0$ (2.30), ignoring the correlation penalty. Once again the simple approximation formula performs well, particularly for extreme values of z, which are likely to be the ones of interest in applications. The correlation penalty is more severe here than in Figure 2, especially in the tails.

The $\mathbf{Cov}_0$ formula (2.30) is essentially the covariance function for a Brownian bridge. Results related to Theorem 1 can be found in the empirical process literature; see equation (2.2) of Csörgő and Mielniczuk (1996) for example, which applies to the "one-class" case when all the z_i's are $\mathcal{N}(0, 1)$. Desai, Deller, and McCormick (2009) extend the covariance calculations in Efron (2007a) to include skewness corrections.

3. ESTIMATION OF THE CORRELATION PARAMETERS

Application of Section 2's theory requires us to estimate several parameters: the rms correlation α (2.22), and the class components (p_c, μ_c, σ_c) in (2.4)–(2.5) (though we will see that the latter task can be avoided under some assumptions). This section illustrates the estimation process in terms of the leukemia study of Section 1. **X**, the data matrix for the study, has $N = 7128$ rows, one for each gene, and $n = 72$ columns, one for each patient; the $n_1 = 47$ ALL patients precede the $n_2 = 25$ AML patients. Entry x_{ij} of **X** is the expression level for gene i

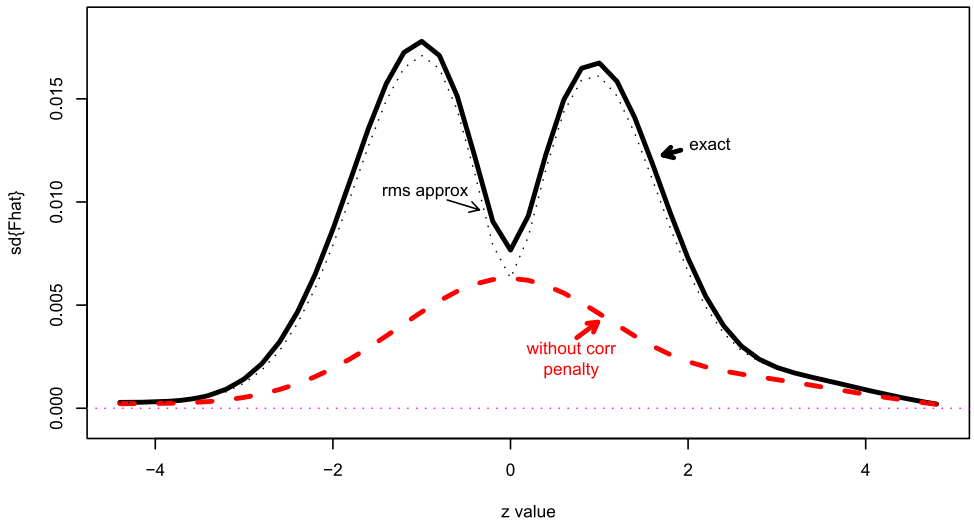

Figure 3. Comparison of exact formula for $\mathrm{sd}\{\hat{F}_k\}$ from Theorem 1 (heavy curve) with Rms approximation using (2.33) (dotted curve); same example as in Figure 2. Dashed curve shows standard deviation estimates ignoring the correlation penalty. The online version of this figure is in color.

on patient j. The columns of $\mathbf{X}$ were individually standardized to have mean 0 and variance 1; see Remark E.

The ith row of $\mathbf{X}$ gives t_i, the two-sample t-statistic comparing expression levels on gene i for AML versus ALL patients. These are converted to z-values $z_i = \Phi^{-1}(F_{70}(t_i))$ (1.2), whose histogram appears in Figure 1. As noted before, the histogram is much wider near its center than a theoretical $\mathcal{N}(0, 1)$ null distribution: analysis using the `locfdr` program described in Efron (2007b, 2008) estimated that proportion $p_0 = 0.93$ of the genes were "null" (i.e., identically distributed for ALL and AML), and that z-values for the null genes followed a $\mathcal{N}(0.09, 1.68^2)$ distribution.

We wish to estimate the rms correlation α (2.22). Let $\mathbf{X}_0$ indicate an $N \times n_0$ subset of $\mathbf{X}$ pertaining to a single population of subjects, for example the 47 ALL patients. There are $N \cdot (N-1)/2$ sample correlations $\hat\rho_{ii'}$ between rows i and i' of $\mathbf{X}_0$. Computing all of these, or a sufficiently large random sample, yields the empirical mean and variance (m, v) of the $\hat\rho$ distribution,

$$\hat\rho \sim (m, v), \tag{3.1}$$

$(m, v) = (0.002, 0.190^2)$ for the ALL patients. As discussed in section 3 of Efron (2009), standardizing the columns of $\mathbf{X}_0$ to have mean 0 forces $m \doteq 0$, and we will assume $m = 0$ in what follows. [This is equivalent to taking $\alpha_1 = 0$ as we did following (2.21).]

The obvious choice $\bar\alpha = v^{1/2}$ tends to greatly overestimate α: each $\hat\rho_{ii'}$ is nearly unbiased for its true correlation $\rho_{ii'}$, a normal-theory approximation for mean and variance being

$$\hat\rho_{ii'} \sim (\rho_{ii'}, (1-\rho_{ii'}^2)^2/(n-3)) \tag{3.2}$$

(Johnson and Kotz 1970), but the considerable variances in (3.2) can greatly broaden the empirical distribution of the $\hat\rho$'s. Two corrected estimates of α are developed in Efron (2009). The simpler correction formula is

$$\hat\alpha^2 = \frac{n_0}{n_0-1}\left(v - \frac{1}{n_0-1}\right) \tag{3.3}$$

based on an identity between the row and column correlations of $\mathbf{X}_0$. The second approach uses an empirical Bayes analysis of the variance term in (3.3) to justify a more elaborate formula,

$$\tilde\alpha^2 = \tilde v - \frac{3}{n_0-5}\tilde v^2 \qquad \left[\tilde v = \frac{(n_0-3)v-1}{n_0-5}\right]. \tag{3.4}$$

The first three columns of Table 2 compare $\hat\alpha$ with $\tilde\alpha$ for $\mathbf{X}_0$ based on the ALL patients, the AML patients, and both. The final column reports mean $\pm$ standard deviation for $\hat\alpha$ and $\tilde\alpha$ in 100 simulations of model (2.24): $N = 6000, n_1 = n_2 = 40$ patients in each class, true $\alpha = 0.10$; see Remark D. The two estimates are effectively linear functions of each other for typical values of v; $\hat\alpha$, the simpler choice, is preferred by the author.

Table 2. Estimates $\hat\alpha$ and $\tilde\alpha$, (3.3) and (3.4), for rms correlation α (2.22) of leukemia data; also 100 simulations of model (2.24), $N = 6000, n_1 = n_2 = 40$, true $\alpha = 0.10$, showing mean $\pm$ standard deviation

	ALL	AML	Both	Simulation
$\hat\alpha$	0.121	0.109	0.114	0.1054 ± 0.0074
$\tilde\alpha$	0.118	0.092	0.113	0.1045 ± 0.0075

It seems that we need to estimate the class components (p_c, μ_c, σ_c) in (2.4)–(2.5) in order to apply the theory of Section 2, but under certain assumptions this can be finessed, as discussed next.

The marginal density $f(z)$ under model (2.4)–(2.5) is

$$f(z) = \sum_c p_c \varphi\left(\frac{z-\mu_c}{\sigma_c}\right)\frac{1}{\sigma_c}; \tag{3.5}$$

so, letting $\mathbf{f} = f(\mathbf{x})$ (the density evaluated at the K-vector of bin midpoints), we have $\Delta \cdot \mathbf{f} = \sum_c p_c \boldsymbol{\pi}_c$ as in (2.8). Formula (2.13) can be expressed as

$$\mathbf{cov}_0 = N\left\{\operatorname{diag}(\Delta\mathbf{f}) - \sum_c p_c \boldsymbol{\pi}_c \boldsymbol{\pi}_c'\right\}. \tag{3.6}$$

Here we are assuming, as in (2.5), that the class sample sizes N_c are fixed. A more realistic assumption might be that the numbers $N_1, N_2, \ldots, N_C$ are a multinomial sample of size N, sampled with probabilities $p_1, p_2, \ldots, p_C$, in which case (3.6) becomes the usual multinomial covariance matrix

$$\mathbf{cov}_0 = N\{\operatorname{diag}(\Delta\mathbf{f}) - \Delta^2 \mathbf{f}\mathbf{f}'\}. \tag{3.7}$$

A smooth curve $\hat{\mathbf{f}}$ fit to the histogram heights [the estimate used here is a Poisson spline regression as described following (4.10)] as in Figure 1 then yields $\widehat{\mathbf{cov}}_0$ by substitution into (3.7), *without requiring knowledge of the class structure* (2.4)–(2.5). In the same way, we can estimate the $\mathbf{Cov}_0$ for $\hat{\mathbf{F}}$ in (2.30) by the standard multinomial formula

$$(\widehat{\mathbf{Cov}}_0)_{kl} = \frac{1}{N}\left\{\hat F_{\max(k,l)} - \hat F_k \hat F_l\right\}. \tag{3.8}$$

Under some circumstances, a similar tactic can be applied to estimate the correlation penalties $\mathbf{cov}_1$ and $\mathbf{Cov}_1$, (2.23) and (2.33). The first and second derivatives $f^{(1)}(z)$ and $f^{(2)}(z)$ of (3.5) are

$$f^{(1)}(z) = \sum_c p_c \varphi^{(1)}\left(\frac{z-\mu_c}{\sigma_c}\right)\frac{1}{\sigma_c^2} \qquad \text{and}$$
$$f^{(2)}(z) = \sum_c p_c \varphi^{(2)}\left(\frac{z-\mu_c}{\sigma_c}\right)\frac{1}{\sigma_c^3}. \tag{3.9}$$

Suppose we make the *homogeneity assumption* that all σ_c values are the same, say $\sigma_c = \sigma_0$. Comparison with definitions (2.20) and (2.32) then gives

$$\bar{\boldsymbol{\varphi}}^{(1)} = \sigma_0^2 \mathbf{f}^{(1)} \qquad \text{and} \qquad \bar{\boldsymbol{\phi}}^{(2)} = \sigma_0^2 \mathbf{f}^{(2)} \tag{3.10}$$

with $\mathbf{f}^{(j)} = (f^{(j)}(x_k))'$. This leads to the convenient covariance penalty formulas,

$$\mathbf{Cov}_1 \doteq \frac{(\sigma_0^2\alpha)^2}{2}\mathbf{f}^{(1)}\mathbf{f}^{(1)\prime} \qquad \text{and}$$
$$\mathbf{cov}_1 \doteq \frac{(N\Delta\sigma_0^2\alpha)^2}{2}\mathbf{f}^{(2)}\mathbf{f}^{(2)\prime} \tag{3.11}$$

from (2.33) and (2.23).

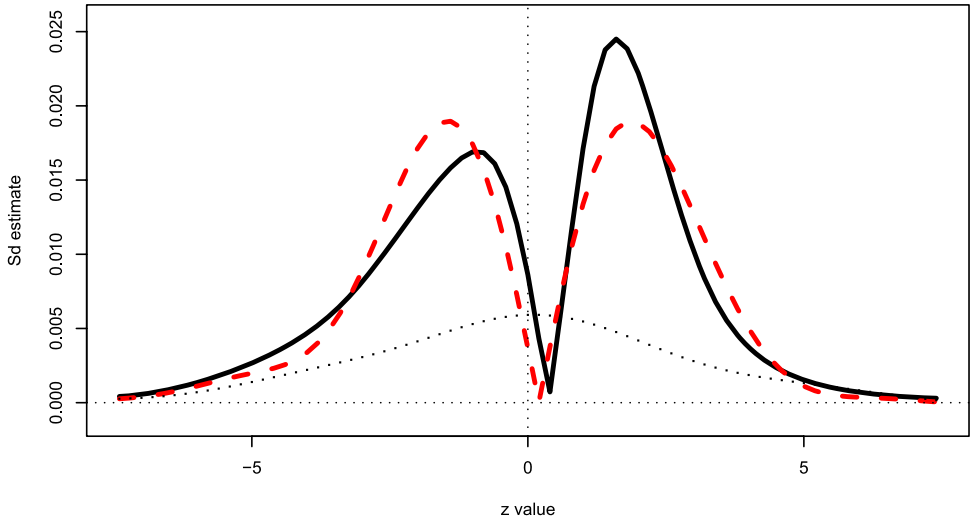

Figure 4. Leukemia data; two estimates of correlation penalty standard deviation $\mathrm{sd}_1\{\hat{F}_k\}$ for $\hat{F}_k$ (2.27). *Solid curve* formula (3.12); *dashed curve* Rms approximation (2.33) using class estimates from Table 3. *Dotted curve* is independence estimate from (3.8), indicating that the correlation penalty is substantial. The online version of this figure is in color.

A smooth estimate $\hat{f}(z)$ of $f(z)$ can be differentiated to give estimated values of $\mathbf{Cov}_1$ and $\mathbf{cov}_1$, for example,

$$\mathrm{sd}_1\{\hat{F}_k\} = (\widehat{\mathbf{Cov}}_1)_{kk}^{1/2} = \frac{\hat{\sigma}_0^2\hat{\alpha}}{\sqrt{2}}\left|\hat{f}^{(1)}(x_k)\right| \tag{3.12}$$

for the correlation penalty standard deviation of $\hat{F}(x_k)$ (2.27). [This provides the second term in formula (1.4).] The heavy curve in Figure 4 shows (3.12) for the leukemia data, using $\hat{\sigma}_0 = 1.68$, $\hat{\alpha} = 0.114$, and $\hat{f}(z)$ from Figure 1.

Suppose we are unwilling to make the homogeneity assumption. A straightforward approach to estimating $\mathbf{Cov}_1$ or $\mathbf{cov}_1$ requires assessments of the parameters (p_c, μ_c, σ_c) in (2.4)–(2.5). These can be based on the "nonnull counts" (Efron 2007b), the small bars plotted negatively in Figure 1; see Remark B. The figure suggests three classes, left, center, and right, with parameter values as estimated in Table 3.

The dashed curve in Figure 4 shows $\mathrm{sd}_1(\hat{F}_k)$ estimated directly from (2.32)–(2.33) using the values in Table 3. It is similar to the homogeneity estimate (3.12) except in the extreme tails.

Formula (1.4) for the standard deviation of $\hat{F}(x)$ was tested in a simulation experiment. The specifications were the same as in Figure 3, with $N = 6000$, $\alpha = 0.10$, and two classes of z-values (2.24). One hundred $\mathbf{X}$ matrices were generated as in the simulation for Table 2, each yielding a vector of 6000 correlated z-values, followed by $\hat{\sigma}_0$, $\hat{\alpha}$, and $\hat{f}^{(1)}(x)$ for use in (1.4); see Remark D for further details. Finally, $\widehat{\mathrm{sd}}$, the square root of (1.4), was calculated along with $\widehat{\mathrm{sd}}_0$, the square root of just the first term.

Table 3. Three-class model (2.4)–(2.5) for leukemia data. Parameter estimates based on nonnull counts, Remark B

	Left	Center	Right
p_c	0.054	0.930	0.016
μ_c	−4.2	0.09	5.4
σ_c	1.16	1.68	1.05

The solid curve in Figure 5 shows the average of the $\widehat{\mathrm{sd}}$ values for x between −4 and 4.5, with solid bars indicating standard deviations of the 100 $\widehat{\mathrm{sd}}$'s. There is a good match of the average with the exact sd curve from Figure 3. The error bars indicate moderate variability across the replications. The average for $\widehat{\mathrm{sd}}_0$, dashed curve, agrees with the corresponding curve in Figure 3 and shows that correlation cannot be ignored in this situation.

4. APPLICATIONS

Correlation usually degrades statistical accuracy, an important question for the data analyst being the severity of its effects on the estimates and tests at hand. The purpose of Sections 2 and 3 was to develop practical methods for honestly assessing the accuracy of inferences made in the presence of large-scale correlation. This section presents a few examples of the methodology in action.

We have already seen one example: in Table 1 the accuracy of $\hat{F}(x)$, the right-sided empirical cdf for the leukemia data, computed from the usual binomial formula that assumes independence among the z-values,

$$\widehat{\mathrm{sd}}_0 = \left\{\hat{F}(x)(1-\hat{F}(x))/N\right\}^{1/2}, \tag{4.1}$$

was compared with $\widehat{\mathrm{sd}}$ from formula (1.4) in which the correlation penalty term was included: $\widehat{\mathrm{sd}}$ more than doubled $\widehat{\mathrm{sd}}_0$ over most of the range.

Suppose we assume, as in Efron (2008), that each of the N cases (the N genes in the leukemia study) is either *null* or *nonnull* with prior probability p_0 or $p_1 = 1 - p_0$, and with the corresponding z-values having density either $f_0(z)$ or $f_1(z)$,

$$\begin{aligned} p_0 &= \Pr\{\text{null}\}, \qquad f_0(z) \text{ density} \quad \text{if null,} \\ p_1 &= \Pr\{\text{nonnull}\}, \qquad f_1(z) \text{ density} \quad \text{if nonnull.} \end{aligned} \tag{4.2}$$

Let F_0 and F_1 be the right-sided cdfs of f_0 and f_1, and F the mixture cdf

$$F(x) = p_0F_0(x) + p_1F_1(x). \tag{4.3}$$

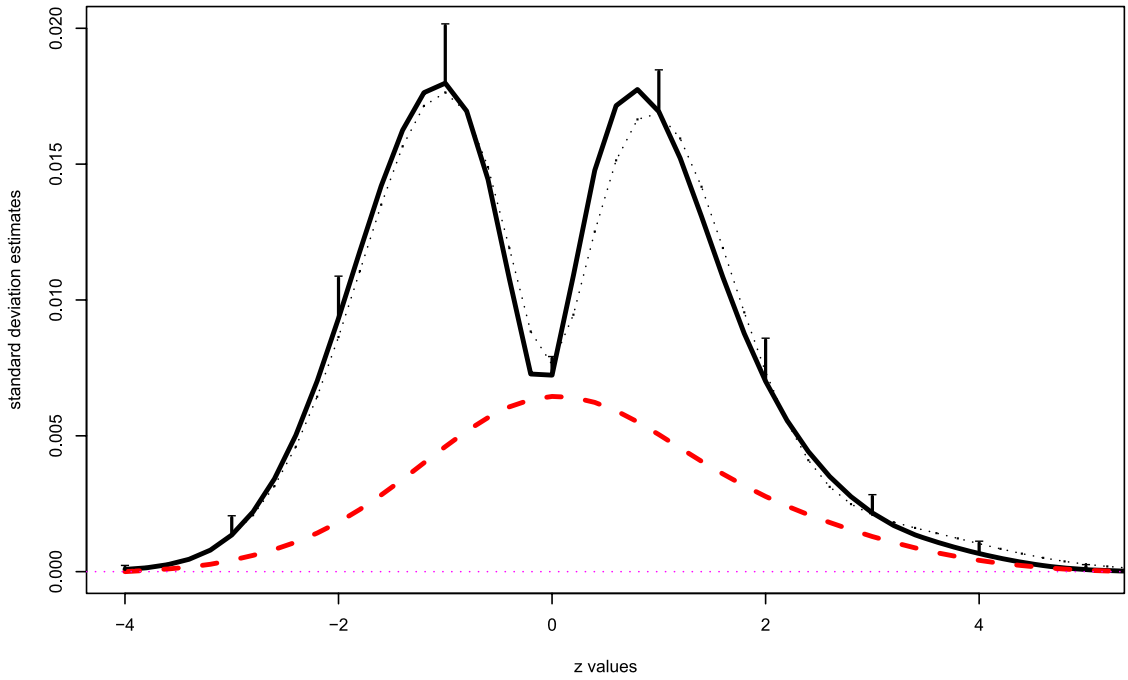

Figure 5. Simulation experiment for formula (1.4). *Solid curve* average of $\widehat{\text{sd}}$, square root of (1.4), 100 replications, with bars indicating standard deviation of $\widehat{\text{sd}}$ at $x = -4, -3, \ldots, 4$; *dotted curve* exact sd from Figure 3; *dashed curve* average of $\widehat{\text{sd}}_0$, standard error estimate for $\hat{F}(x)$ ignoring correlation. The online version of this figure is in color.

The probability of a case being null given that z exceeds x is

$$\text{Fdr}(x) \equiv \Pr\{\text{null}|z \geq x\} = \frac{p_0 F_0(x)}{F(x)} \tag{4.4}$$

according to Bayes theorem, "fdr," standing for false discovery rate.

If p_0 and F_0 are known then fdr has the obvious estimate

$$\widehat{\text{Fdr}}(x) = p_0 F_0(x)/\hat{F}(x) \tag{4.5}$$

(2.1). Benjamini and Hochberg's celebrated 1995 algorithm uses $\widehat{\text{Fdr}}(x)$ for simultaneous hypothesis testing, but it can also be thought of as an empirical Bayes estimator of the Bayesian probability $\text{Fdr}(x)$. The bottom row of Table 1 shows $\widehat{\text{Fdr}}(x)$ for the leukemia data, taking $p_0 = 0.93$ and $F_0 \sim \mathcal{N}(0.09, 1.68^2)$ as in Figure 1. (Later we will do a more ambitious calculation taking into account the estimation of p_0 and F_0.)

The coefficient of variation for $\widehat{\text{Fdr}}(x)$ approximately equals that for $\hat{F}(x)$ [when $p_0 F_0(x)$ is known in (4.5)]. At $x = 5$ we have $\widehat{\text{Fdr}}(5) = 0.15$, with coefficient of variation about 0.19. An $\widehat{\text{Fdr}}$ of 0.15 might be considered small enough to trigger significance in the Benjamini–Hochberg algorithm, but in any case it seems clear that the probabiltiy of being null is quite low for the 71 genes having z_i above 5. Even taking account of correlation effects, we have a rough upper confidence limit of 0.21 [i.e., $0.15 \cdot (1 + 2 \cdot 0.19)$] for Fdr(5).

Next we consider accuracy estimates for a general class of statistics $Q(\mathbf{y})$, where Q is a q-dimensional function of the count vector $\mathbf{y}$ (2.3). As in section 5 of Efron (2007b), we assume that a small change $d\mathbf{y}$ in the count vector (considered as varying continuously) produces change dQ in Q according to

$$dQ = \hat{\mathbf{D}}\, d\mathbf{y} \qquad [\hat{D}_{jk} = \partial Q_j/\partial y_k]. \tag{4.6}$$

If $\widehat{\mathbf{cov}}(\mathbf{y})$ is a covariance estimate for $\mathbf{y}$, obtained perhaps as in (2.12), (3.8), or (3.11), then the usual delta-method estimate for $\mathbf{cov}(Q)$ is

$$\widehat{\mathbf{cov}}(Q) = \hat{\mathbf{D}}\widehat{\mathbf{cov}}(\mathbf{y})\hat{\mathbf{D}}'. \tag{4.7}$$

In a theoretical context, where $\mathbf{cov}(\mathbf{y})$ is known, we might instead use

$$\text{cov}(Q) \doteq \mathbf{D}\mathbf{cov}(\mathbf{y})\mathbf{D}' \tag{4.8}$$

now with the derivative matrix $\mathbf{D}$ evaluated at the expectation of $\mathbf{y}$.

Model (4.2) yields the *local false discovery rate*

$$\text{fdr}(x) \equiv \Pr\{\text{null}|z = x\} = p_0 f_0(x)/f(x), \tag{4.9}$$

$f(x)$ being the mixture density

$$f(x) = p_0 f_0(x) + p_1 f_1(x); \tag{4.10}$$

$\text{fdr}(x)$ is inferentially more appropriate than $\text{Fdr}(x)$ from a Bayesian point of view, but it is not as immediately available since it involves estimating the *density* $f(x)$. However, because z-value densities are mixtures of near-normals as shown in Section 5, it is usually straightforward to carry out the estimation.

`Locfdr`, the algorithm discussed in Efron (2007a, 2008), estimates $f(x)$ by means of Poisson regression of the counts y_k as a spline function of the x_k, the bin midpoints in (2.2)–(2.3). The structure matrix $\mathbf{M}$ for the Poisson regression is $K \times d$, where K is the number of bins and d is degrees of freedom (e.g., the number of free parameters of the spline fit; see Remark A for details). Let $\hat{\mathbf{f}}$ be the vector of fitted values $\hat{f}(x_k)$, and $\hat{\boldsymbol{\ell}}$ the vector with components $\hat{\ell}_k = \log(\hat{f}(x_k))$. Then, as discussed in Efron (2007b, section 5), (4.6) takes the form

$$d\hat{\boldsymbol{\ell}} = \hat{\mathbf{D}}\, d\mathbf{y} \quad \text{with } \hat{\mathbf{D}} = \mathbf{M}(\mathbf{M}'\,\text{diag}(N\Delta\hat{\mathbf{f}})\mathbf{M})^{-1}\mathbf{M}' \tag{4.11}$$

and we can use (4.7) or (4.8) to approximate $\mathbf{cov}(\hat{\boldsymbol{\ell}})$.

For any function $v(x)$ define the vector

$$\mathbf{v} = (v_1, v_2, \ldots, v_k, \ldots, v_K)' = (\ldots, v(x_k), \ldots)' \tag{4.12}$$

as with $\hat{\mathbf{f}}$ and $\hat{\boldsymbol{\ell}}$ above. If

$$\widehat{\mathrm{lfdr}}(x) \equiv \log(\widehat{\mathrm{fdr}}(x)) = \log(p_0 f_0(x)) - \log(f(x)) \tag{4.13}$$

then

$$\widehat{\mathbf{lfdr}}(x) = \log(p_0) + \log(\mathbf{f}_0) - \hat{\boldsymbol{\ell}} \tag{4.14}$$

implying, if p_0 and f_0 are known, that

$$\mathbf{cov}(\widehat{\mathbf{lfdr}}(x)) = \mathbf{cov}(\hat{\boldsymbol{\ell}}) \doteq \mathbf{D}\mathbf{cov}(\mathbf{y})\mathbf{D}' \tag{4.15}$$

with $\mathbf{D} = \mathbf{M}(\mathbf{M}'\,\mathrm{diag}(N\Delta\mathbf{f})\mathbf{M})^{-1}\mathbf{M}'$ (4.8).

The solid curves in Figure 6 plot standard deviations for $\log(\widehat{\mathrm{fdr}}(x))$, obtained as square root of the diagonal elements of $\mathbf{cov}(\widehat{\mathbf{lfdr}})$ (4.15), for model (2.24) with $N = 6000$ and rms correlation α equal 0, 0.1, or 0.2; see Remark D. The horizontal axes are plotted in terms of the upper percentiles of $F(x)$, the right end of each plot corresponding to the far right tail of the z-value distribution. For $\alpha = 0$, $\mathrm{sd}(\log\widehat{\mathrm{fdr}}(x))$ increases from 0.03 to 0.08 as we move from the fifth to the first percentile of F. The coefficient of variation (CV) of $\widehat{\mathrm{fdr}}(x)$ nearly equals $\mathrm{sd}(\log\widehat{\mathrm{fdr}}(x))$, so $\widehat{\mathrm{fdr}}(x)$ is quite accurately estimated for $\alpha = 0$, but substantially less so for $\alpha = 0.2$. Reducing N to 1500 doubles the standard deviation estimates for $\alpha = 0$, but has less effect in the correlated situations: for $\alpha = 0.1$ for example, the increase is only 20% at percentile 0.025. Simulations confirmed the correctness of these results.

Intuitively it seems that fdr should be harder to estimate than Fdr, but that is not what Figure 6 shows. Let $\hat{L}_k = \log(\hat{F}(x_k))$, with corresponding vector $\hat{\mathbf{L}}$. Then $\hat{\mathbf{D}}$ in (4.6) has

$$\hat{D}_{jk} = B_{jk}/(N \cdot \hat{F}_j) \tag{4.16}$$

with $\mathbf{B}$ as in (2.25), giving an estimate of $\mathbf{cov}(\hat{\mathbf{L}})$ from (4.7) or (4.8). The same argument as (4.13)–(4.15) shows that this also estimates $\mathbf{cov}(\log\widehat{\mathrm{Fdr}})$, the log of vector (4.5), assuming $p_0F_0(x)$ is known. The dotted curves in Figure 6 show standard deviations for $\log(\widehat{\mathrm{Fdr}}(x))$. If anything, Figure 6 suggests that $\widehat{\mathrm{fdr}}$ is *less* variable than $\widehat{\mathrm{Fdr}}$, particularly at the smaller percentiles.

Here we are comparing the nonparametric estimator $\widehat{\mathrm{Fdr}}(x)$ (4.5) with the parametric estimator $\widehat{\mathrm{fdr}}(x)$. The Poisson spline estimate $\hat{f}(x)$ that gave $\widehat{\mathrm{fdr}}(x)$ can be summed to give parametric estimates of $F(x)$ and $\mathrm{Fdr}(x)$, say $\widetilde{\mathrm{Fdr}}(x)$. Straightforward calculations show that the derivative matrix $\hat{\mathbf{D}}$ for $\widetilde{\mathrm{Fdr}}(x)$ is

$$\hat{\mathbf{D}} = \hat{\mathbf{C}}\hat{\mathbf{D}}_f, \quad \text{where } C_{jk} = B_{jk}\hat{f}_k/\hat{F}_j \tag{4.17}$$

with $\mathbf{B}$ from (2.25) and $\hat{\mathbf{D}}_f$ equaling $\hat{\mathbf{D}}$ in (4.11). Standard deviations for $\log(\widetilde{\mathrm{Fdr}})$, shown by the dashed curves in Figure 6, indicate about the same accuracy for $\widetilde{\mathrm{Fdr}}(x)$ as for $\widehat{\mathrm{Fdr}}$.

All of these calculations assumed that p_0 and $f_0(z)$ [or $F_0(z)$] in (4.2) were known. This is unrealistic in situations like the leukemia study, where there is clear evidence that a textbook $\mathcal{N}(0, 1)$ theoretical null distribution is too narrow. Estimating an "empirical null" distribution, such as $\mathcal{N}(0.09, 1.68^2)$ in Figure 1, is both necessary and feasible (see Efron 2008) but can greatly increase variability, as discussed next.

Formula (4.14) becomes

$$\widehat{\mathbf{lfdr}} = \log(\hat{p}_0) + \log(\hat{\mathbf{f}}_0) - \hat{\boldsymbol{\ell}} \tag{4.18}$$

when p_0 and f_0 are themselves estimated. The corresponding derivative matrix $\hat{\mathbf{D}} = d\widehat{\mathbf{lfdr}}/d\mathbf{y}$ in (4.6) appears as equation (5.8) in Efron (2007b), this formula applying to the *central matching* method for estimating $p_0f_0(z)$. The second row of Table 4 shows $\mathrm{sd}\{\log\widehat{\mathrm{fdr}}(x)\}$ obtained from $\mathbf{Dcov}(\mathbf{y})\mathbf{D}'$ for the same situation as in the middle panel of Figure 6. Comparison with the theoretical null standard deviations (from the

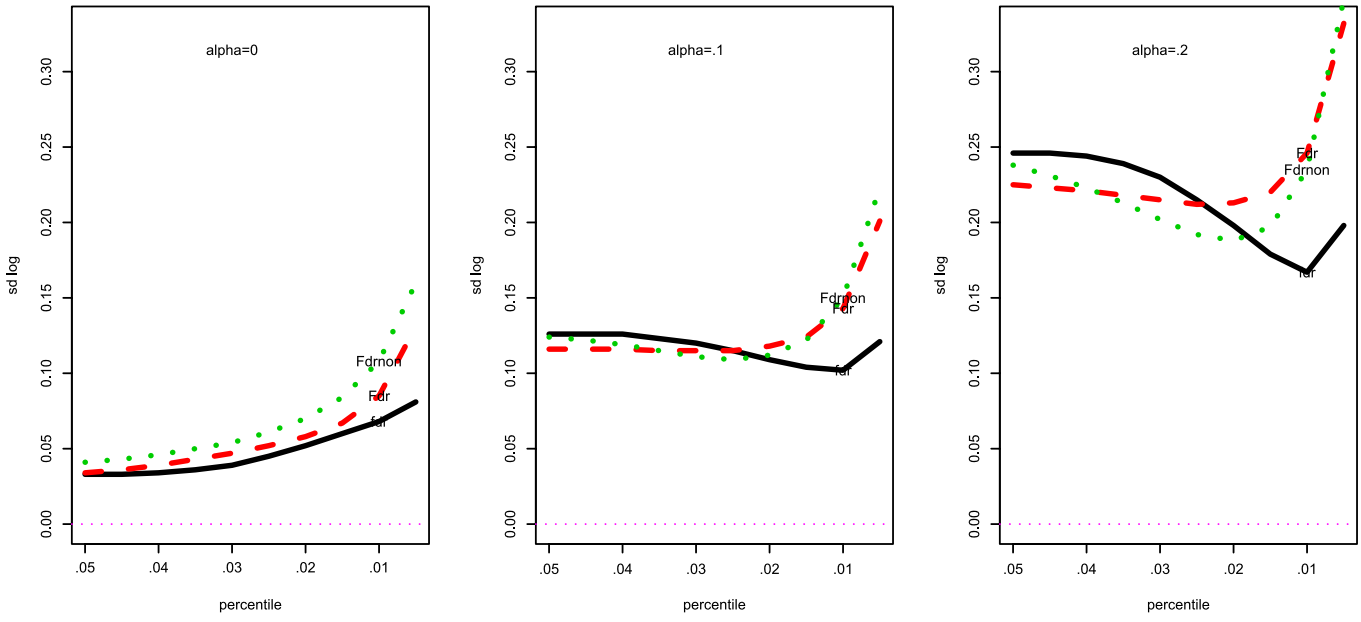

Figure 6. *Solid curves* show standard deviation of $\log(\widehat{\mathrm{fdr}}(x))$ as a function of x at the upper percentiles of the z-value distribution for model (2.24), $N = 6000$ and $\alpha = 0, 0.1, 0.2$. *Dotted curves* (green) same for $\log(\widehat{\mathrm{Fdr}}(x))$ (4.5), nonparametric Fdr estimator. *Dashed curves* (red) for parametric version (4.17) of Fdr estimator. The online version of this figure is in color.

Table 4. Comparison of sd{log($\widehat{\text{fdr}}(x)$)} using empirical null versus theoretical null for the situation in the middle panel of Figure 6. The empirical null standard deviations are much larger, as seen also in Efron (2007b)

	Percentile				
	0.05	0.04	0.03	0.02	0.01
sd empirical null	**0.18**	**0.26**	**0.36**	**0.54**	**0.83**
sd theoretical null	0.13	0.13	0.12	0.11	0.10
x	1.98	2.16	2.40	2.74	3.25
fdr(x)	0.69	0.58	0.44	0.25	0.09
Fdr(x)	0.34	0.27	0.19	0.10	0.04

solid curve in the middle panel) shows that estimating the null distribution greatly increases variability.

Here are some points to note:

- Accuracy is worse for log($\widehat{\text{Fdr}}$) then for log($\widehat{\text{fdr}}$) in the top line of Table 4.
- Accuracy is somewhat better when $p_0 f_0(z)$ is estimated by the MLE option in `locfdr` (lemma 2 of Efron 2007b).
- The big empirical null standard deviations in Table 4 are at least partially misleading: some of the variability in $\widehat{\text{fdr}}(x)$ is "signal" rather than "noise," tracking conditional changes in the appropriate value of fdr(x). See figure 2 of Efron (2007a) and the discussion in that paper.

Remark F of Section 6 describes a parametric bootstrap resampling scheme that avoids the Taylor series computations of (4.7), but which has not yet been carefully investigated.

5. THE NONNULL DISTRIBUTION OF z-VALUES

The results of the previous sections depend on the variates z_i having normal distributions (1.5). By definition, a z-value is a statistic having a $\mathcal{N}(0, 1)$ distribution under a null hypothesis H_0 of interest (1.1): but will it still be normal for nonnull conditions? This section shows that under repeated sampling the nonnull distribution of z will typically have mean $O(1)$, standard deviation $1 + O(n^{-1/2})$, and nonnormality $O_p(n^{-1})$ (as measured by the magnitude of skewness and kurtosis). In other words, normality degrades more slowly than unit standard deviation as we move away from the null hypothesis.

Figure 7 illustrates the phenomenon for the case of noncentral t distributions,

$$z = \Phi^{-1}(F_\nu(t)), \qquad t \sim t_\nu(\delta), \tag{5.1}$$

the notation indicating a noncentral t variable with ν degrees of freedom and noncentrality parameter δ (*not* δ^2), as described in chapter 31 of Johnson and Kotz (1970). Here, as in (1.2), F_ν is the cdf of a *central* t_ν distribution. The standard deviation of z decreases as $|\delta|$ increases; for $\delta = 5$, $\nu = 20$, z has (mean, sd) equal $(4.01, 0.71)$. The useful and perhaps surprising observation is that normality holds up quite well even far from the null case $\delta = 0$. We tacitly used this fact to justify application of our theoretical results to the leukemia study.

To begin the theoretical development, suppose that $y_1, y_2, \ldots, y_n$ are independent and identically distributed (iid) observations sampled from F_θ, a member of a one-parameter family of distributions,

$$\mathcal{F} = \{F_\theta, \theta \in \Theta\} \tag{5.2}$$

having its moment parameters {mean, standard deviation, skewness, kurtosis}, denoted

$$\{\mu_\theta, \sigma_\theta, \gamma_\theta, \delta_\theta\}, \tag{5.3}$$

defined differentiably in θ. The results that follow are heuristic in the sense that they only demonstrate second-order Cornish–Fisher expansion properties, with no attempt to provide strict error bounds.

Under the null hypothesis $H_0 : \theta = 0$, which we can write as

$$H_0 : y \sim \{\mu_0, \sigma_0, \gamma_0, \delta_0\}, \tag{5.4}$$

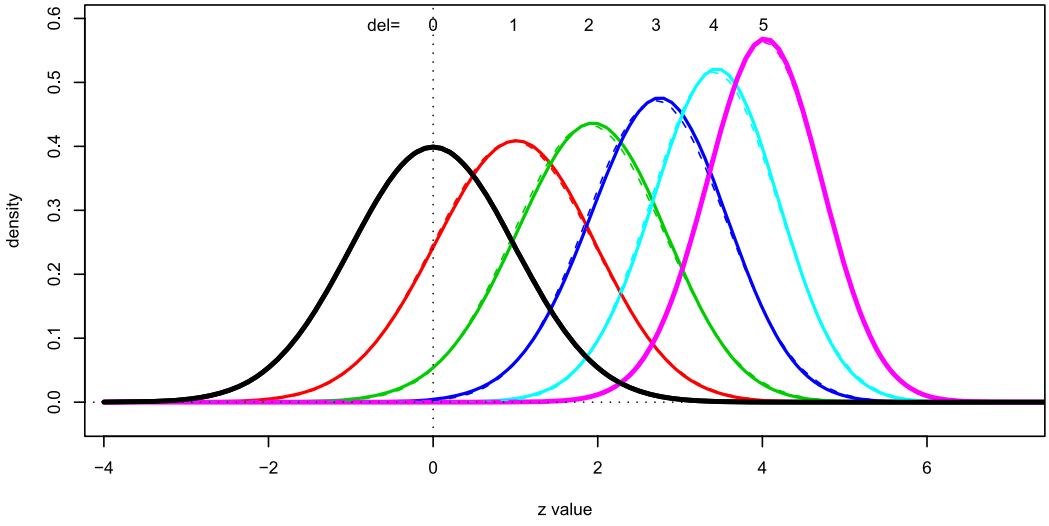

Figure 7. Density of the z-value statistic (5.1) when t has a noncentral t distribution with $\nu = 20$ degrees of freedom; for noncentrality parameter $\delta = 0, 1, 2, 3, 4, 5$. The densities are seen to be nearly normal; dashed curves are exact normal densities matched in mean and standard deviation. For $\delta = 5$, z has (mean, sd, skew, kurt) $= (4.01, 0.71, -0.06, 0.08)$. Negative values of δ give mirror image results. Remark G of Section 6 describes the density function calculations. The online version of this figure is in color.

the standardized variate

$$Y_0 = \sqrt{n}\left(\frac{\bar{y}-\mu_0}{\sigma_0}\right) \qquad \left[\bar{y} = \sum_{i=1}^{n} y_i/n\right] \tag{5.5}$$

satisfies

$$H_0 : Y_0 \sim \left\{0, 1, \frac{\gamma_0}{\sqrt{n}}, \frac{\delta_0}{n}\right\}. \tag{5.6}$$

Normality can be improved to second order by means of a Cornish–Fisher transformation,

$$Z_0 = Y_0 - \frac{\gamma_0}{6\sqrt{n}}(Y_0^2 - 1) \tag{5.7}$$

which reduces the skewness in (5.6) from $O(n^{-1/2})$ to $O(n^{-1})$,

$$H_0 : Z_0 \sim \{0, 1, 0, 0\} + O(n^{-1}). \tag{5.8}$$

See chapter 1 of Johnson and Kotz (1970) or, for much greater detail, section 2.2 of Hall (1992). We can interpret (5.8) as saying that Z_0 is a *second-order z-value*,

$$H_0 : Z_0 \sim \mathcal{N}(0, 1) + O_p(n^{-1}), \tag{5.9}$$

e.g., *a test statistic giving standard normal p-values accurate to* $O(n^{-1})$.

Suppose now that H_0 is false, and instead H_1 is true, with $y_1, y_2, \ldots, y_n$ iid according to

$$H_1 : y \sim \{\mu_1, \sigma_1, \gamma_1, \delta_1\} \tag{5.10}$$

rather than (5.4). Setting

$$Y_1 = \sqrt{n}\left(\frac{\bar{y}-\mu_1}{\sigma_1}\right) \qquad \text{and} \qquad Z_1 = Y_1 - \frac{\gamma_1}{6\sqrt{n}}(Y_1^2 - 1) \tag{5.11}$$

makes Z_1 second-order normal under H_1,

$$H_1 : Z_1 \sim \mathcal{N}(0, 1) + O_p(n^{-1}). \tag{5.12}$$

We wish to calculate the distribution of Z_0 (5.7) under H_1. Define

$$c = \sigma_1/\sigma_0, \qquad d = \sqrt{n}(\mu_1 - \mu_0)/\sigma_0, \qquad \text{and} \qquad g_0 = \gamma_0/(6\sqrt{n}). \tag{5.13}$$

Some simple algebra yields the following relationship between Z_0 and Z_1.

Lemma 3. Under definitions (5.7), (5.11), and (5.13),

$$Z_0 = M + SZ_1 + g_0\left\{\left(\frac{\gamma_1}{\gamma_0}S - c^2\right)(Y_1^2 - 1) + (1 - c^2)\right\}, \tag{5.14}$$

where

$$M = d \cdot (1 - dg_0) \qquad \text{and} \qquad S = c \cdot (1 - 2dg_0). \tag{5.15}$$

The asymptotic relationships claimed at the start of this section are easily derived from Lemma 3. We consider a sequence of alteratives θ_n approaching the null hypothesis value θ_0 at rate $n^{-1/2}$,

$$\theta_n - \theta_0 = O\left(n^{-1/2}\right). \tag{5.16}$$

The parameter $d = \sqrt{n}(\mu_{\theta_n} - \mu_0)/\sigma_0$ defined in (5.13) is then of order $O(1)$, as is

$$M = d(1 - dg_0) = d(1 - d\gamma_0/(6\sqrt{n})), \tag{5.17}$$

while standard Taylor series calculations give

$$c = 1 + \frac{\dot{\sigma}_0}{\dot{\mu}_0}\frac{d}{\sqrt{n}} + O(n^{-1}) \qquad \text{and} \qquad S = 1 + \left(\frac{\dot{\sigma}_0}{\dot{\mu}_0} - \frac{\gamma_0}{3}\right)\frac{d}{\sqrt{n}} + O(n^{-1}), \tag{5.18}$$

the dot indicating differentiation with respect to θ.

Theorem 2. Under model (5.2), (5.16), and the assumptions of Lemma 3,

$$Z_0 \sim \mathcal{N}(M, S^2) + O_p(n^{-1}) \tag{5.19}$$

with M and S as given in (5.17)–(5.18). Moreover,

$$\left.\frac{dS}{dM}\right|_{\theta_0} = \frac{1}{\sqrt{n}}\left(\left.\frac{d\sigma}{d\mu}\right|_{\theta_0} - \frac{\gamma_0}{3}\right) + O(n^{-1}). \tag{5.20}$$

Proof. The proof of Theorem 2 uses Lemma 3, with θ_n playing the role of H_1 in (5.14). Both $1 - c^2$ and $(\gamma_1/\gamma_0)S - c^2$ are of order $O(n^{-1/2})$; the former from (5.18) and the latter using $\gamma_1/\gamma_0 = 1 + (\dot{\gamma}_0/\gamma_0)(\theta_n - \theta_0) + O(n^{-1})$. Since $Y_1^2 - 1$ is $O_p(1)$, this makes the bracketed term in (5.14) $O_p(n^{-1/2})$; multiplying by $g_0 = \gamma_0/(6\sqrt{n})$ reduces it to $O_p(n^{-1})$, and (5.19) follows from (5.12). Differentiating M and S in (5.17)–(5.18) with respect to d verifies (5.20).

Theorem 2 supports our claim that, under nonnull alternatives, the null hypothesis normality of Z_0 degrades more slowly than its unit standard deviation, the comparison being $O_p(n^{-1})$ versus $O(n^{-1/2})$.

One-parameter exponential families are an important special case of (5.2). With θ the natural parameter of $\mathcal{F}$ and y its sufficient statistic, that is, with densities proportional to $\exp\{\theta y\}g_0(y)$, (5.20) reduces to

$$\left.\frac{dS}{dM}\right|_{\theta_0} = \frac{\gamma_0}{6\sqrt{n}} + O(n^{-1}). \tag{5.21}$$

The parameter $\gamma_0/(6\sqrt{n})$ is called the *acceleration* in Efron (1987), interpreted as "the rate of change of standard deviation with respect to expectation on the normalized scale," which agrees with its role in (5.21).

As an example, suppose $y_1, y_2, \ldots, y_n \overset{\text{iid}}{\sim} \theta\Gamma_1$, Γ_n indicating a standard gamma distribution with n degrees of freedom, so $g_\theta(y) = (1/\theta)\exp(-y/\theta)$ for $y \geq 0$. (Equivalently, $\bar{y} \sim \theta\Gamma_n/n$.) This is an exponential family having skewness $\gamma_0 = 2$ for any choice of θ_0. An exact z-value for testing $H_0 : \theta = \theta_0$ is

$$Z_0 = \Phi^{-1}(G_n(n\bar{y}/\theta_0)), \tag{5.22}$$

where G_n is the cdf of Γ_n. Table 5 shows the mean, standard deviation, skewness, and kurtosis of Z_0 for $n = 10$, $\theta_0 = 1$, evaluated for several choices of the alternative θ_1. The standard deviation of Z_0 increases steadily with θ_1; here $\gamma_0/(6\sqrt{n}) = 0.1054$, matching to better than three decimal places the observed numerical derivative dS/dM. Skewness and kurtosis are both very

Table 5. Gamma example, $n = 10, \theta_0 = 1$, indicating the distribution of z-value (5.22) for various nonnull choices of θ. Standard deviation increases with θ_1 in accordance with (5.21), while maintaining near-perfect normality for Z_0

	θ_1						
	0.4	0.5	0.67	1	1.5	2.0	2.5
mean	−2.49	−1.94	−1.19	0	1.36	2.45	3.38
stdev	0.76	0.81	0.88	1	1.15	1.27	1.38
skew	−0.05	−0.04	−0.02	0	0.02	0.04	0.04
kurt	0.01	0.01	0.00	0	0.00	−0.01	−0.04

small; in the equivalent of Figure 7, there is no visible discrepancy at all between the density curves for Z_0 and their matching normal equivalents.

So far we have considered z-values obtained from an average $\bar{y}$ of iid observations, but the results of Theorem 2 hold in greater generality. Section 5 of Efron (1987) considers one-parameter families where $\hat{\theta}$, an estimator of θ, has MLE-like asymptotic properties in terms of its bias, standard deviation, skewness, and kurtosis,

$$\hat{\theta} \sim \{\theta + \beta_\theta/n, \sigma_\theta/\sqrt{n}, \gamma_\theta/\sqrt{n}, \delta_\theta/n\}. \tag{5.23}$$

Letting $\hat{\theta}$ play the role of $\bar{y}$ and $\mu_\theta = \theta + \beta_\theta/n$ in definitions (5.5)–(5.12), Lemma 3, and Theorem 2 remain true, assuming only the validity of the Cornish–Fisher transformations (5.9)–(5.12). Ignoring the bias β_θ, that is, taking $Y_0 = \sqrt{n}(\hat{\theta} - \theta_0)/\sigma_0$ at (5.5), adds an $O(n^{-1/2})$ term to M in (5.17).

Moving beyond one-parameter families, suppose $\mathcal{F}$ is a p-parameter exponential family, having densities proportional to $\exp\{\eta_1 x_1 + \eta_2' x_2\} g_0(x_1, x_2)$, where η_1 and x_1 are real-valued while η_2 and x_2 are $(p-1)$-dimensional vectors, but where we are only interested in η_1, not the nuisance vector η_2. The *conditional* distribution of x_1 given x_2 is then a one-parameter exponential family with natural parameter η_1, which puts us back in the context of Theorem 2. Remark H of Section 6 suggests a further extension where the parameter of interest "θ" can be a general real-valued function of η, not just a coordinate such as η_1.

The noncentral t family does not meet the conditions of Lemma 3 or Theorem 2: (5.1) is symmetric in δ around zero, causing γ_0 in (5.14) to equal zero and likewise the derivative in (5.20). Nevertheless, as Figure 7 shows, it does exhibit impressive nonnull normality. Table 6 displays the moment parameters of $z = \Phi^{-1}(F_\nu(t))$ (1.3), for $t \sim t_\nu(\delta)$, $\nu = 20$ and $\delta = 0, 1, 2, 3, 4, 5$. The nonnull normality is not quite as good as

Table 6. Noncentral t example $t \sim t_\nu(\delta)$ for $\nu = 20$, $\delta = 0, 1, 2, 3, 4, 5$; moment parameters of $z = \Phi^{-1}(F_\nu(t))$ (1.3) indicate near-normality even for δ far from 0. [Moments calculated using (6.10)]

	δ					
	0	1	2	3	4	5
mean	0	0.98	1.89	2.71	3.41	4.01
sd	1	0.98	0.92	0.85	0.77	0.71
skew	0	−0.07	−0.11	−0.11	−0.10	−0.07
kurt	0	0.02	0.06	0.08	0.09	0.07

in the gamma example of Table 5, but is still quite satisfactory for its application in Section 4.

Microarray studies can be more elaborate than two-sample comparisons. Suppose that in addition to the $N \times n$ expression matrix $\mathbf{X}$ we have measured a primary response variable y_j and covariates $w_{j1}, w_{j2}, \ldots, w_{jp}$ on each of the n subjects. Given the observed expression levels $x_{i1}, x_{i2}, \ldots, x_{in}$ for gene i, we could calculate t_i, the usual t-value for y_j as a function of x_{ij}, in a linear model that includes the p covariates. Then

$$z_i = \Phi^{-1}(F_{n-p-1}(t_i)) \tag{5.24}$$

is a z-value (1.1) under the usual Gaussian assumption, showing behavior like that in Table 6 for nonnull genes.

6. REMARKS

Some remarks, proofs, and details relating to the previous sections are presented here.

Remark A (Poisson regression). The curve $\hat{f}(z)$ in Figure 1 is a Poisson regression fit to the counts y_k, as a natural spline function of the bin centers x_k. Here the x_k ranged from −7.8 to 7.8 in steps of $\Delta = 0.2$, while the spline had five degrees of freedom, so $\mathbf{M}$ in (4.11) was 79×6 (including the intercept column). See section 5 of Efron (2007b).

Remark B (Table 3). Section 3 of Efron (2008) defines the *nonnull counts* $y_k^{(1)} = (1 - \widehat{\text{fdr}}(x_k)) \cdot y_k$. Since, under model (4.2), $1 - \widehat{\text{fdr}}(x_k)$ estimates the proportion of nonnull z-values in bin k, $y_k^{(1)}$ estimates the number of nonnulls. The $y_k^{(1)}$ values are plotted below the x axis in Figure 1, determining the "left" and "right" distribution parameters in Table 3. "Center" was determined by the empirical null fit from `locfdr`, using the MLE method described in section 4 of Efron (2007b). This method tends to underestimate the nonnull counts near $z = 0$, and also the σ_c values for the left and right classes, but increasing then to 1.68 had little effect on the dashed curve in Figure 4.

Remark C (Proof of Lemma 1). Let $I_k(i)$ denote the indicator function of the event $z_i \in \mathcal{Z}_k$ (2.2) so that the number of z_i's from class $\mathcal{C}_c$ in $\mathcal{Z}_k$ is

$$y_{kc} = \sum_{\mathbf{c}} I_k(i), \tag{6.1}$$

the boldface subscript indicating summation over the members of $\mathcal{C}_c$. We first compute $E\{y_{kc} y_{ld}\}$ for bins k and l, $k \neq l$, and classes c and d,

$$\begin{aligned} E\{y_{kc} y_{ld}\} &= E\Big\{\sum_{\mathbf{c}}\sum_{\mathbf{d}} I_k(i) I_l(j)\Big\} \\ &= \Delta^2 \sum_{\mathbf{c}}\sum_{\mathbf{d}} \varphi_{\rho_{ij}}(x_{kc}, x_{ld})(1 - \chi_{ij})/(\sigma_c \sigma_d) \end{aligned} \tag{6.2}$$

following notation (2.5)–(2.10), with χ_{ij} the indicator function of event $i = j$ (which can only occur if $c = d$). This reduces to

$$\begin{aligned} E\{y_{kc} y_{ld}\} &= N^2 \Delta^2 p_c (p_d - \chi_{cd}/N) \\ &\quad \times \int_{-1}^{1} \varphi_\rho(x_{kc}, x_{ld}) g(\rho)\, d\rho/(\sigma_c \sigma_d) \end{aligned} \tag{6.3}$$

under the assumption that the same correlation distribution $g(\rho)$ applies across all class combinations. Since $y_k = \sum_c y_{kc}$ (2.3), we obtain

$$E\{y_k y_l\} = N^2\Delta^2 \sum_c \sum_d p_c(p_d - \chi_{cd}/N) \times \int_{-1}^{1} \varphi_\rho(x_{kc}, x_{ld}) g(\rho)\, d\rho/(\sigma_c\sigma_d), \quad (6.4)$$

the nonbold subscripts indicating summation over classes.

Subtracting

$$E\{y_k\}E\{y_l\} = N^2\Delta^2 \sum_c \sum_d \varphi(x_{kc})\varphi(x_{ld})/(\sigma_c\sigma_d) \quad (6.5)$$

from (6.4) results, after some rearrangement, in

$$\begin{aligned}\operatorname{cov}(y_k, y_l) = N^2\Delta^2 &\sum_c \sum_d \frac{\varphi(x_{kc})\varphi(x_{ld})}{\sigma_c\sigma_d} \\ &\times \left\{ p_c\left(p_d - \frac{\chi_{cd}}{N}\right) \right. \\ &\left. \times \int_{-1}^{1} \left(\frac{\varphi_\rho(x_{kc}, x_{ld})}{\varphi(x_{kc})\varphi(x_{ld})} - 1 \right) g(\rho)\, d\rho \right\} \\ &- N\Delta^2 \sum_c p_c \frac{\varphi(x_{kc})\varphi(x_{ld})}{\sigma_c\sigma_d}. \end{aligned} \quad (6.6)$$

Using $\pi_{kc} = \Delta \cdot \varphi(x_{kc})/\sigma_c$ as in (2.8), expression (6.6) is seen to equal the klth element of $\mathbf{cov}(\mathbf{y})$ in Lemma 1, when $k \neq l$.

The case $k = l$ proceeds in the same way, the only difference being that $N\Delta p_c \chi_{cd}\varphi(x_{kc})/\sigma_c$ must be added to formula (6.3). This adds $N\Delta \sum_c p_c\varphi(x_{kc})/\sigma_c$ to (6.6), again in agreement with $\mathbf{cov}(\mathbf{y})$ in Lemma 1.

The assumption that $g(\rho)$ is the same across all classes can be weakened for the rms approximations (2.23) and (2.33), where we only need the second moments α_2 to be the same. In fact, the class structure can disappear entirely for rms formulas, as seen in (3.12).

Remark D [Model (2.24)]. Specifications (2.24) were recentered to give overall expectation 0 in (2.4), (2.5):

$$(p_0, \mu_0, \sigma_0) = (0.95, -0.125, 1) \quad \text{and} \quad (p_1, \mu_1, \sigma_1) = (0.05, 2.38, 1), \quad (6.7)$$

these being the parameter values used in Figures 2, 3, 5, and 6. Recentering overall expectations to zero is common in practice, a consequence of the data matrix $\mathbf{X}$ having its column-wise means subtracted off.

The 6000×80 data matrices $\mathbf{X}$ used in the simulations for Table 2 and Figure 5 had entries $x_{ij} \sim \mathcal{N}(\delta_{ij}, 1)$ independent across columns j: $\delta_{ij} = 0$ for $j \le 40$, while for columns $j > 40$,

$$\delta_{ij} = 0.224\mu_1 \quad \text{for } i = 1, 2, \ldots, 300 \quad \text{and} \quad \delta_{ij} = 0.224\mu_0 \quad \text{for } i > 300; \quad (6.8)$$

z-values based on the difference of means between the last and first 40 "patients" then satisfy (2.4), (6.7). The correlation distribution $g(\rho)$ was supported on two points, 20% $\rho = 0.20$ and 80% $\rho = -0.05$, giving $\alpha = 0.10$.

Remark E (Leukemia data standardization). The original entries x_{ij} of the leukemia data matrix were genetic expression levels obtained using Affymetrix oligonucleotide microarrays. For the analyses here, each column of $\mathbf{X}$ was replaced by its normal score values $\tilde{x}_{ij} = \Phi^{-1}((r_{ij} - 0.5)/7128)$, where r_{ij} was the rank of x_{ij} in its column. Transformations such as this reduce the disturbing effects of sensitivity differences between microarrays; see Bolstad et al. (2003).

Remark F (A parametric bootstrap method). Section 3 of Efron (2007a) discusses a hierarchical Poisson simulation scheme that can be adapted to the more general context of this paper. Following the notation in (3.11), we first simulate a vector $\mathbf{u}$,

$$\mathbf{u} = N\Delta\left(\hat{\mathbf{f}} + \frac{\sigma_0^2}{\sqrt{2}} A\hat{\mathbf{f}}^{(2)}\right) \quad \text{with } A \sim \mathcal{N}(0, \hat{\alpha}^2), \quad (6.9)$$

and then take $\mathbf{y} \sim \text{Poisson}(\mathbf{u})$, that is $y_k \overset{\text{ind}}{\sim} \text{Poisson}(u_k)$. The simulated $\mathbf{y}$ vectors can then be used to assess the variability of any function $Q(\mathbf{y})$, obviating the need for the derivative matrix $\hat{\mathbf{D}}$. This amounts to a parametric bootstrap approach to accuracy estimation. It produced similar answers to (4.11) when applied to the leukemia data but seemed prone to biases in other applications. (Nonparametric bootstrapping, resampling columns of the data matrix $\mathbf{X}$, can produce erratic results for the kind of large-scale accuracy problems considered in Section 4.)

Remark G (z-value densities). Suppose test statistic t has possible densities $\{f_\theta(t), \theta \in \Theta\}$, with corresponding cdfs $F_\theta(t)$, and we wish to test $H_0 : \theta = \theta_0$. The z-value statistic $z = \Phi^{-1}\{F_{\theta_0}(t)\}$ then has densities

$$g_\theta(z) = \varphi(z) f_\theta(t)/f_{\theta_0}(t). \quad (6.10)$$

The density curves in Figure 7 were obtained from (6.10), with $f_\theta(t)$ the noncentral $t_\nu(\theta)$ density, $\nu = 20$ and $\theta = 0, 1, 2, 3, 4, 5$.

Remark H (Extensions of Theorem 2). In some circumstances, Theorem 2 can be extended to multiparameter families $\mathcal{F} = \{F_\eta\}$ where we wish to test $\theta = \theta_0$ for θ a real-valued function of η. This is straightforward to verify in the context of Efron (1985), which includes for example Fieller's problem, and is conjectured to be true in general exponential families.

7. SUMMARY

The paper considers studies where a large number N of cases are under investigation, N perhaps in the hundreds or thousands, each represented by its own z-value z_i, and where there is the possibility of substantial correlation among the z_i's. Our main result is a simple approximation formula for the accuracy of summary statistics such as the empirical cdf of the z-values or an estimated false discovery rate. The argument proceeds in five steps:

- Exact formulas for the accuracy of correlated z-value cdfs are derived under normal distribution assumptions (Section 2).
- Simple approximations to the exact formulas are developed in terms of the root mean square correlation of all $N \cdot (N-1)/2$ cases [Section 2, (2.23), and (2.33)].

- Practical estimates for the approximation formulas are derived and demonstrated through simulations and application to a microarray study (Sections 3 and 4).
- Delta-method arguments are used to extend the cdf results to more general summary statistics (Sections 3 and 4).
- Under reasonable assumptions, it is shown that z scores tend to have nearly normal distributions, even in nonnull situations (Section 5), justifying application of the theory to studies in which the individual variates are z-values.

Our main conclusion is that by dealing with normal variates, a practical assessment of large-scale correlation effects on statistical estimates is possible.

[Received March 2009. Revised June 2009.]

REFERENCES

Bolstad, B., Irizarry, R., Astrand, M., and Speed, T. (2003), "A Comparison of Normalization Methods for High Density Oligonucleotide Array Data Based on Variance and Bias," *Bioinformatics*, 19, 185–193. [1054]

Clarke, S., and Hall, P. (2009), "Robustness of Multiple Testing Procedures Against Dependence," *The Annals of Statistics*, 37, 332–358. [1043]

Csörgő, S., and Mielniczuk, J. (1996), "The Empirical Process of a Short-Range Dependent Stationary Sequence Under Gaussian Subordination," *Probability Theory and Related Fields*, 104, 15–25. [1046]

Desai, K., Deller, J., and McCormick, J. (2009), "The Distribution of Number of False Discoveries for Highly Correlated Null Hypotheses," *The Annals of Applied Statistics*, to appear. [1043,1046]

Dudoit, S., van der Laan, M. J., and Pollard, K. S. (2004), "Multiple Testing. I. Single-Step Procedures for Control of General Type I Error Rates," *Statistical Applications in Genetics Molecular Biology*, 3, Article 13 (electronic). [1043]

Dudoit, S., Shaffer, J. P., and Boldrick, J. C. (2003), "Multiple Hypothesis Testing in Microarray Experiments," *Statistical Science*, 18, 71–103. [1042]

Efron, B. (1985), "Bootstrap Confidence Intervals for a Class of Parametric Problems," *Biometrika*, 72, 45–58. [1054]

——— (1987), "Better Bootstrap Confidence Intervals" (with discussion), *Journal of the American Statistical Association*, 82, 171–200. [1052,1053]

——— (2007a), "Correlation and Large-Scale Simultaneous Significance Testing," *Journal of the American Statistical Association*, 102, 93–103. [1042-1046,1049,1051,1054]

——— (2007b), "Size, Power and False Discovery Rates," *The Annals of Statistics*, 35, 1351–1377. [1047-1051,1053]

——— (2008), "Microarrays, Empirical Bayes and the Two-Groups Model" (with discussion), *Statistical Science*, 23, 1–22. [1047-1050,1053]

——— (2009), "Are a Set of Microarrays Independent of Each Other?" *The Annals of Applied Statistics*, 3, 922–942. [1045,1047]

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., and Lander, E. S. (1999), "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring," *Science*, 286, 531–537. DOI: *10.1126/science.286.5439.531*. [1042,1043]

Hall, P. (1992), *The Bootstrap and Edgeworth Expansion. Springer Series in Statistics*, New York: Springer-Verlag. [1052]

Johnson, N. L., and Kotz, S. (1970), *Distributions in Statistics. Continuous Univariate Distributions*, Vol. 1, Boston, MA: Houghton Mifflin. [1047,1051,1052]

Lancaster, H. O. (1958), "The Structure of Bivariate Distributions," *The Annals of Mathematical Statistics*, 29, 719–736. [1044]

Owen, A. B. (2005), "Variance of the Number of False Discoveries," *Journal of the Royal Statistical Society, Ser. B*, 67, 411–426. [1042,1043]

Qiu, X., Brooks, A., Klebanov, L., and Yakovlev, A. (2005), "The Effects of Normalization on the Correlation Structure of Microarray Data," *BMC Bioinformatics*, 6, 120. DOI: *10.1186/1471-2105-6-120*. [1043]

Qiu, X., Klebanov, L., and Yakovlev, A. (2005), "Correlation Between Gene Expression Levels and Limitations of the Empirical Bayes Methodology for Finding Differentially Expressed Genes," *Statistical Applications in Genetics and Molecular Biology*, 4, Article 34 (electronic). [1043]

Schwartzman, A., and Lin, X. (2009), "The Effect of Correlation in False Discovery Rate Estimation," Biostatistics Working Paper 106, Harvard University. [1043]

Westfall, P., and Young, S. (1993), *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment*, New York: Wiley-Interscience. [1042]

Comment

T. Tony Cai

Professor Efron has given us an interesting article on the effects of correlation which is an important issue in multiple testing. He is to be congratulated for his significant contributions to large-scale multiple testing.

Much of the research on multiple testing has been focused on the independent case and many practical testing procedures have been developed under the independence assumption. However, in many interesting applications observations are correlated. It is known that correlation has significant effects on a multiple testing procedure. For example, both the expectation and variance of the number of Type I errors can be seriously affected by the correlation among the test statistics (see Finner and Roters 2002 and Owen 2005). Correlation can also substantially deteriorate the performance of many FDR procedures. Previous research on the effects of correlation in large-scale multiple testing has been mostly focused on the validity of various testing procedures under dependency. For example, Benjamini and Yekutieli (2001), Farcomeni (2007), and Wu (2009) show that the FDR is controlled at the nominal level by the BH step-up and adaptive p-value procedures under different dependence assumptions.

Among various aspects of the correlation effects on an FDR procedure, the validity issue is often overemphasized. FDR procedures developed under the independence assumption, even valid, may suffer from substantial efficiency loss when the dependence structure is ignored. These situations include the geographical disease mapping studies, multiple-stage clinical trials, functional Magnetic Resonance Imaging analyses and comparative microarray experiments, where the nonnull cases are often structured in some way, for example, correlated temporally, spatially, or functionally.

A critical step in the implementation of many FDR procedures is the estimation of several important quantities such as the proportion of the nonnull hypotheses, the empirical null distribution, and the true FDR level. Developing good estimators

T. Tony Cai is Professor, Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104 (E-mail: *tcai@wharton.upenn.edu*). Research supported in part by NSF FRG grant DMS-0854973.

DOI: 10.1198/jasa.2010.tm10241

for these quantities is a challenging task, even in the independent case.

The present paper continues a prolific line of research of Professor Efron on multiple testing, but with a different focus from his previous papers. Efron (2007) considered the correlation effects on the null distribution of the z-values and suggested that an adjusted FDR estimate should be combined with the use of an Lfdr procedure to remove the bias caused by the correlation. The present paper focuses on the effects of correlation on certain summary statistics used in large-scale multiple testing procedures. It is demonstrated that when the test statistics are correlated, the distribution of the z-values and its functionals can be conveniently estimated with adjustments for correlation without the need of knowing the whole covariance matrix, which itself is very difficult, if not impossible, to estimate in the current setting. A key step of the proposed estimation procedure is the use of a clever approximation based on Mehler's identity and the root mean square correlation. The approximation can be efficiently carried out, which makes the method practical for applications.

The convenient correlation adjustments proposed in the paper are obtained through a sequence of approximations. It is illustrated in the paper that the method works well in several examples. It is of great interest to further study the precision of these estimators and to understand how well the targeted quantities can be optimally estimated under reasonable assumptions. Perhaps even more important questions are how to best use these estimators to construct more efficient multiple testing procedures under dependency and how the performance of these estimators affects the performance of the subsequent testing procedures.

As mentioned earlier, the estimation step plays an important role in a multiple testing procedure. Several alternative estimation methods have been developed in the literature. For example, in the independent case, Cai and Jin (2010) developed a frequency domain approach based on the empirical characteristic function and Fourier analysis for the estimation of both the parameters associated with the null distribution f_0 and the proportion of the nonnull effects p_0. These estimators were shown to attain the optimal rates of convergence under regularity conditions. In the correlated case, the estimators were shown to be uniformly consistent over a wide class of parameters when the dependency is short ranged or strongly mixing (see Jin and Cai 2007). Numerical results also showed that the estimators perform favorably in comparison to other existing methods.

It is true that, as pointed out in the paper, "correlation usually degrades statistical accuracy, an important question for the data analyst being the severity of its effects on the estimates and tests at hand." However, in many settings correlation effects can also be positive on the outcomes of a testing procedure. Intuitively it is clear that the dependency structure among hypotheses is highly informative in simultaneous inference and can be exploited to construct more efficient tests. For example, in comparative microarray experiments, it is found that changes in expression for genes can be the consequence of regional duplications or deletions, and significant genes tend to appear in clusters. Therefore, when deciding the significance level of a particular gene, the observations from its neighborhood should also be taken into account.

It is possible to construct significantly better multiple testing procedures in the correlated settings by modeling the dependency structure. Sun and Cai (2009) considered multiple testing under a particular dependency structure, the hidden Markov model (HMM). The HMM is an effective tool for modeling the dependency structure and has been widely used in areas such as speech recognition, signal processing as well as analysis biological sequences and processes. Using a compound decision theoretical framework, an oracle testing procedure is developed in an ideal setting where the HMM parameters are assumed to be known. Under mild conditions, the oracle procedure is shown to be optimal in the sense that it minimizes the false nondiscovery rate (FNR) subject to a constraint on the FDR. This approach is distinguished from the conventional methods in that the proposed procedure is built on a new test statistic (local index of significance, LIS) instead of the p-values. Unlike p-values, the LIS takes into account the observations in adjacent locations by exploiting the local dependency structure in the HMM. The precision of individual tests is hence improved by utilizing the dependency information.

A data-driven procedure is then constructed to mimics the oracle procedure by plugging in consistent estimates of the unknown HMM parameters. The data-driven procedure is shown to be asymptotically optimal in the sense that it attains both the FDR and FNR levels of the oracle procedure asymptotically. Numerical studies indicate the favorable performance of the LIS procedure. These findings show that the correlation among hypotheses can be highly informative in large-scale simultaneous inference and can be exploited to construct more efficient testing procedures.

Much research is still needed in order to fully understand the correlation effects on the accuracy of estimators used in multiple testing procedures as well as the testing procedures themselves under general dependency structures. The present paper raises important questions and will definitely stimulate new research in the future.

ADDITIONAL REFERENCES

Benjamini, Y., and Yekutieli, D. (2001), "The Control of False Discovery Rate in Multiple Testing Under Dependency," *The Annals of Statistics*, 29, 1165–1188. [1055]

Cai, T., and Jin, J. (2010), "Optimal Rates of Convergence for Estimating the Null and Proportion of Non-Null Effects in Large-Scale Multiple Testing," *The Annals of Statistics*, 38, 100–145. [1056]

Farcomeni, A. (2007), "Some Results on the Control of the False Discovery Rate Under Dependence," *Scandinavian Journal of Statistics*, 34, 275–297. [1055]

Finner, H., and Roters, M. (2002), "Multiple Hypotheses Testing and Expected Number of Type I Errors," *The Annals of Statistics*, 30, 220–238. [1055]

Jin, J., and Cai, T. (2007), "Estimating the Null and the Proportion of Non-Null Effects in Large-Scale Multiple Comparisons," *Journal of the Amererican Statistical Association*, 102, 495–506. [1056]

Sun, W., and Cai, T. (2009), "Large-Scale Multiple Testing Under Dependence," *Journal of the Royal Statistical Society, Ser. B*, 71, 393–424. [1056]

Wu, W. (2009), "On False Discovery Control Under Dependence," *The Annals of Statistics*, 36, 364–380. [1055]

Comment

Ruth HELLER

Professor Efron has given us an interesting article on how to quantify the uncertainty in summary statistics of interest in large-scale problems, when the summary statistics are based on correlated normal variates. It is shown that the inflation in the accuracy estimate due to correlation among the normal variates cannot be ignored (except possibly at the very far tails of distributions).

Using a series of simplifications of the covariance formula, a simple formula is derived and it is shown in a numerical example that the approximation is indeed very close to the truth. In particular it is shown that the entire correlation structure is captured by one parameter α, the rms correlation. Several methods of estimating α, as well as the other unknown parameters, are suggested.

In what follows I will discuss several topics in large scale significance testing that are related to the results of this paper.

SUMMARY VALUES OF INTEREST WHEN *z*–VALUES ARE CORRELATED

In large-scale significance testing, methods that control or estimate the false discovery rate are often applied to identify the set of interesting discoveries. Professor Efron suggested the following two summary statistics and estimated their accuracy: the estimated tail-area false discovery rate $\widehat{\mathrm{Fdr}}(x) = p_0 F_0(x)/\hat{\mathrm{F}}(x)$ and the estimated local false discovery rate $\widehat{\mathrm{fdr}}(x) = p_0 f_0(x)/\hat{f}(x)$. Keeping the same notations as in the manuscript, $F_0(\cdot)$ and $\hat{\mathrm{F}}(\cdot)$ are the null and the empirical survival curve of the z-values.

Another summary statistics of interest is the following quantity that "monotonizes" the $\widehat{\mathrm{Fdr}}(z)$ curve: $\overline{\mathrm{Fdr}}(z) = \inf\{z' \le z : \widehat{\mathrm{Fdr}}(z')\}$. In practice, the cutoff value x is often chosen to be the smallest z-value so that $\widehat{\mathrm{Fdr}}(z) \le q$, where q is a desired fraction chosen to be typically small (e.g., $q = 0.05, 0.1, 0.25$), and the hypotheses with z-values above this cutoff x are reported as interesting discoveries. The same cutoff value x is selected when choosing the smallest z-value so that $\overline{\mathrm{Fdr}}(z) \le q$. This practice coincides with the celebrated Benjamini and Hochberg's false discovery rate controlling procedure (Benjamini and Hochberg 1995), henceforth referred to as the BH procedure, up to the factor p_0, conservatively taken as 1.

The BH procedure appears to control the false discovery rate in most circumstances that are not highly artificial (Romano, Shaikh, and Wolf 2008; Yekutieli 2008). When the test statistics are correlated, the false discovery proportion (FDP, the fraction of discoveries from null hypotheses out of all discoveries) of the specific dataset at hand may be very high even though the false discovery rate, $\mathrm{FDR} = E(\mathrm{FDP})$, is controlled at the nominal level q on average over (hypothetical) replications of the study (Pawitan, Calza, and Ploner 2006; Efron 2007a). For a given dataset, the interest of the investigator is in the FDP, not the FDR. When the rms correlation is nonnegligible, the FDP may be much higher than q.

Similarly, for a given dataset and a given cutoff value x, the interest is in the false discovery proportion $\mathrm{FDP}(x)$, the fraction of z-scores above x from null hypotheses out of all z-scores above x. The variance of $\mathrm{FDP}(x)$ depends critically on the correlation among the z-values. When the correlation is weak and the effect sizes are large, the variability of $\mathrm{FDP}(x)$ is tight around its expectation, $\mathrm{FDR}(x) = E(\mathrm{FDP}(x))$. In this favorable setting there may be interest in the quantities $\mathrm{FDR}(x)$, $\mathrm{Fdr}(x) = p_0 F_0(x)/F(x)$ or $\mathrm{fdr}(x) = p_0 f_0(x)/f(x)$. However, when the correlation is high or the effect sizes are small, interest may no longer be in $\mathrm{FDR}(x)$, $\mathrm{Fdr}(x)$ or $\mathrm{fdr}(x)$ but in the unknown random quantity $\mathrm{FDP}(x)$. The estimated $\mathrm{Fdr}(x)$ (and its variability) may therefore not be of interest when the estimated rms correlation is nonnegligible. However, from Efron (2007a) and Pawitan, Calza, and Ploner (2006) it appears that using the histogram of z-scores, it may be possible to identify whether the $\mathrm{FDP}(x)$ is indeed much higher than expected for a given dataset.

A SIMULATION STUDY TO ILLUSTRATE THE EFFECT OF CORRELATION ON THE FDP

Each of 1000 datasets was generated as follows: $N = 1000$ genes with expression values from two classes with parameters $(p_0, \mu_0, \sigma_0) = (0.95, 0, 1)$ and $(p_1, \mu_1, \sigma_1) = (0.05, 1, 1)$; 40 cases were generated each from $X_i \sim \mathrm{MVN}(\vec{0}, \Sigma)$ and 40 cases were generated each from $X_i \sim \mathrm{MVN}(\vec{\mu}, \Sigma)$; the first 50 entries in $\vec{\mu}$ are one and the remaining 950 entries are zero; Σ is a block diagonal correlation matrix, each block of size 200 with symmetric correlation of 0.5. The data matrix $\mathbf{X}$ was either standardized by subtracting off its column-wise means (as done in Professor Efron's paper) or left unstandardized. The correlation in each dataset was substantial: the rms correlation was $\alpha = 0.2$. For comparison, 1000 datasets were also generated under independence (i.e., Σ was the identity matrix).

The z-score for row i was $z_i = \Phi^{-1}(F_{78}(t_i))$, where $t_i = (\sum_{j=41}^{80} x_{ij}/40 - \sum_{j=1}^{40} x_{ij}/40)/\widehat{\mathrm{SE}}$ is the t-statistic for comparing the mean of the last 40 cases with that of the first 40 cases, and F_{78} and Φ are the cumulative distribution functions for a Student-t distribution with 78 degrees of freedom and a standard normal respectively. The BH procedure was applied at nominal level $q = 0.01, 0.02, \ldots, 0.25$ to the z-scores in each dataset. Figure 1 top shows the 50th, 75th, and 95th quantiles of the FDP for each q after applying the BH procedure to each

Ruth Heller is Senior Lecturer, Faculty of Industrial Engineering and Management, Technion—Israel Institute of Technology, Haifa, Israel (E-mail: *ruheller@ techunix.technion.ac.il*).

Journal of the American Statistical Association
September 2010, Vol. 105, No. 491, Theory and Methods
DOI: 10.1198/jasa.2010.tm10240

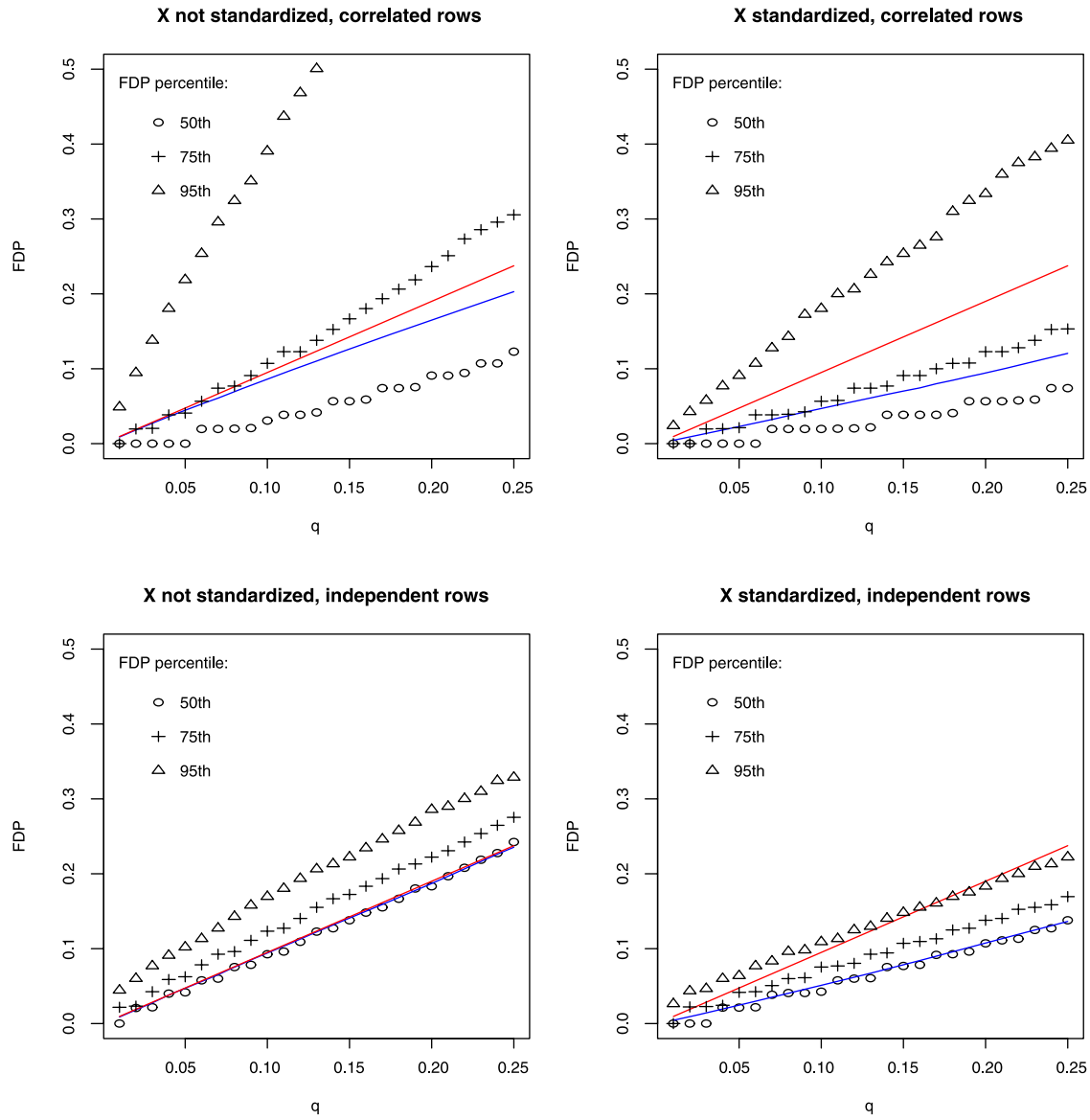

Figure 1. A plot of the 50th, 75th, and 95th FDP percentiles for each level q after applying the BH procedure to each of 1000 simulated datasets. Solid red line is the nominal level $0.95 * q$, blue line is the average FDP.

of 1000 simulated datasets when the data matrix was not standardized (left) and when it was standardized (right). For comparison, the same analysis was repeated when the data was independent and the results are displayed in Figure 1 bottom. The blue line in Figure 1 is the average FDP, the red line is the nominal FDR level $0.95 * q$. The variability of the FDP was much higher when the data was correlated than when it was independent. The standardization of X reduced the variability of the FDP. The variability in the correlated case was reduced as the nominal value q decreased. The average FDP was below the nominal level for all q as expected, and moreover it was almost the same as the nominal level when X was not standardized and the data was independent. When the data was standardized yet independent, the average FDP was below the nominal level since after standardization the p-values were no longer independent, nor were they uniformly distributed but had instead a distribution that was stochastically larger than the uniform.

For each dataset, FDP(x) and $\widehat{\text{Fdr}}(x)$ were computed for the following cutoff values: $x = 2.00, 2.25, 2.50, 2.75, 3.00, 3.50, 4.00, 5.00, 6.00$. Table 1 shows summary statistics of FDP(x) and $\widehat{\text{Fdr}}(x)$ for the correlated case as well as for the independent case (the columns of the data matrix X were not standardized). The average and standard deviation are summarized. For FDP(x), which may be highly skewed in the correlated case, the 50th, 85th, and 95th quantiles are also summarized. While the average FDP(x) was below the average $\widehat{\text{Fdr}}(x)$ for the correlated case, the variability of FDP(x) was very large for the smaller x's and diminished as x increased. For the independence case, the average FDP(x) was very close to the average $\widehat{\text{Fdr}}(x)$, and the variability of FDP(x) was much smaller compared with the correlated case. For example, the effect of correlation cannot be ignored at $x = 2.25$: for correlated data, the average FDP(2.25) was 0.15, but the 85th and 95th quantiles of FDP(2.25) were, respectively, 0.31 and 0.48; for independent data, the average FDP(2.25) was 0.19, and the 85th and 95th quantiles were, respectively, 0.23 and 0.26.

Table 1. Summary statistics of FDP(x) and $\widehat{\text{Fdr}}(x)$ for the correlated case as well as for the independent case (the columns of the data matrix X were not standardized)

	Correlated rows		Independent rows	
x	FDP(x) mean (SD), $Q_{0.5}, Q_{0.85}, Q_{0.95}$	$\widehat{\text{Fdr}}(x)$ mean (SD)	FDP(x) mean (SD), $Q_{0.5}, Q_{0.85}, Q_{0.95}$	$\widehat{\text{Fdr}}(x)$ mean (SD)
2.00	0.24 (0.18), 0.21, 0.45, 0.61	0.33 (0.08)	0.30 (0.04), 0.30, 0.34, 0.37	0.31 (0.02)
2.25	0.15 (0.15), 0.10, 0.31, 0.48	0.20 (0.04)	0.19 (0.05), 0.19, 0.23, 0.26	0.19 (0.01)
2.50	0.09 (0.12), 0.04, 0.18, 0.33	0.11 (0.02)	0.11 (0.04), 0.11, 0.14, 0.17	0.11 (0.01)
2.75	0.05 (0.09), 0.02, 0.09, 0.21	0.06 (0.01)	0.05 (0.03), 0.06, 0.09, 0.11	0.06 (0.00)
3.00	0.02 (0.06), 0.00, 0.04, 0.12	0.03 (0.01)	0.03 (0.02), 0.02, 0.04, 0.06	0.03 (0.00)
3.50	0.01 (0.03), 0.00, 0.00, 0.04	0.01 (0.01)	0.00 (0.01), 0.00, 0.02, 0.03	0.01 (0.00)
4.00	0.00 (0.01), 0.00, 0.00, 0.00	0.00 (0.00)	0.00 (0.01), 0.00, 0.00, 0.00	0.00 (0.00)

NOTE: The average and standard deviation are summarized. For FDP(x), which may be highly skewed in the correlated case, the 50th, 85th, and 95th quantiles are also summarized.

CORRELATION IN RELATION TO SIGNAL

In Professor Efron's paper it is assumed that the correlation is not informative of where the signal lies. However, this assumption may not always apply. For example, different genes may cluster into groups that participate in the same molecular functions or biological process and exhibit high correlation. If these groups are known a-priori, this knowledge can be incorporated into the multiple comparisons procedure to gain power (e.g., Benjamini and Heller 2007; Heller et al. 2009). Aggregates of statistical estimates within each group (and their accuracy) can be useful in this setting. Incorporating the correlation structure without a-priori knowledge of the grouping is a greater challenge [Sun and Cai (2009) model the unknown correlation structure assuming a hidden Markov model for the hypotheses].

In closing, I congratulate Professor Efron for the interesting article, and I thank the editor for giving me an opportunity to contribute to the discussion.

ADDITIONAL REFERENCES

Benjamini, Y., and Heller, R. (2007), "False Discovery Rates for Spatial Signals," *Journal of the American Statistical Association*, 102 (480), 1272–1281. [1059]

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 (1), 289–300. [1057]

Heller, R., Manduchi, E., Grant, G., and Ewens, W. (2009), "A Flexible Two Stage Procedure for Identifying Gene Sets That Are Differentially Expressed," *Bioinformatics*, 25 (8), 1019–1025. [1059]

Pawitan, Y., Calza, S., and Ploner, A. (2006), "Estimation of False Discovery Proportion Under General Dependence," *Bioinformatics*, 22 (24), 3025–3031. [1057]

Romano, J., Shaikh, A., and Wolf, M. (2008), Rejoinder on "Control of the False Discovery Rate Under Dependence Using the Bootstrap and Subsampling," by J. Romano, A. Shaikh, and M. Wolf, *Test*, 17 (3), 461–471. [1057]

Sun, W., and Cai, T. (2009), "Large-Scale Multiple Testing Under Dependence," *Journal of the Royal Statistical Society, Ser. B*, 71 (2), 393–424. [1059]

Yekutieli, D. (2008), Comment on "Control of the False Discovery Rate Under Dependence Using the Bootstrap and Subsampling," by J. Romano, A. Shaikh, and M. Wolf, *Test*, 17 (3), 458–460. [1057]

Comment

Armin SCHWARTZMAN

1. INTRODUCTION

In a series of recent articles, Bradley Efron has pointed out that in large-scale multiple testing problems, the observed distribution of the test statistics often does not match the theoretical null distribution (Efron 2004, 2007a, 2007b, 2008). The correction, which he termed "empirical null," has been a subject of controversy in the statistical community. In Efron (2007a), he made the case that, even when the theoretical model is correct, the observed distribution of the test statistics can look different from the theoretical null distribution simply because of correlation between the test statistics.

Efron's present article represents an important step forward in the understanding of this problem. As opposed to previous papers, where the the effect of correlation was treated within the context of multiple testing problems and false discovery rates, Efron's present article breaks through the confusion by separating these two concepts and focusing on the core issue, which is the behavior of a large collection of correlated normal variables. Only then, as applications, he presents the implications for false discovery rate analysis when the correlated normal variables are z-scores in a large-scale multiple testing problem. I think this separation is crucial and helps get us nearer a new theory of inference for high-dimensional data.

In what follows, I present my own interpretation of Efron's results on how correlation affects the empirical distribution of

Armin Schwartzman is Assistant Professor, Department of Biostatistics, Harvard School of Public Health and Dana–Farber Cancer Institute, Boston, MA 02115 (E-mail: *armins@hsph.harvard.edu*). This work was partially supported by NIH grant PO1-CA134294. The author thanks Rebecca Betensky and Sarah Emerson for helpful discussions.

Journal of the American Statistical Association
September 2010, Vol. 105, No. 491, Theory and Methods
DOI: 10.1198/jasa.2010.tm10237

normal variables. As a shortcut, I work in the countinous domain directly rather than with histogram bins and avoid the inclusion of unnecessary constraints such as normalization, which applies very specifically to microarray data. For simplicity, I assume all the variables are standard normal rather than belonging to a mixture model, but may have an arbitrary correlation structure.

Using these results, I consider the question raised in Efron (2007a) of whether large-scale correlation can substantially widen the observed histogram, as in Figure 1 in that paper and Efron's current one. The theoretical arguments below indicate that the observed histogram is more likely to be narrow than wide, and that it cannot be too wide before it becomes bimodal. An important implication is that a wide unimodal histogram may be an indication of the presence of true signal, rather than an artifact of correlation.

I will conclude commenting briefly on the possibility of performing this kind of analysis with χ^2 variates rather than normal.

I will try to keep the notation close to that of Efron, but some discrepancies in notation will be inevitable.

2. THE DISTRIBUTION OF CORRELATED NORMAL VARIATES

Let $Z_1, Z_2, \ldots, Z_N$ be N(0, 1) variables with pairwise correlations $\rho_{ii'} = \operatorname{cor}(Z_i, Z_{i'})$. Using a similar notation to that in (2.1) of Efron's article, let $\hat{F}(x)$ denote the right-sided empirical cdf

$$\hat{F}(x) = \frac{1}{N}\sum_{i=1}^{N} 1(Z_i \geq x),$$

where $1(\cdot)$ denotes the indicator function. The empirical cdf $\hat{F}(x)$ is always an unbiased estimator of the marginal right-sided standard normal cdf

$$E[\hat{F}(x)] = \Phi^+(x) = 1 - \Phi(x).$$

Its covariance function is given by the following proposition. The result resembles Theorem 1 of Efron, but is given instead in continuous form. The proof is given in the Appendix.

Proposition 1. For any $x, x' \in \mathbb{R}$,

$$\operatorname{cov}[\hat{F}(x), \hat{F}(x')] = \mathbf{Cov}_0(x, x') + \mathbf{Cov}_1(x, x'),$$

where

$$\mathbf{Cov}_0(x, x') = \frac{1}{N}\big[\Phi^+(\max(x, x')) - \Phi^+(x)\Phi^+(x')\big] \quad (1)$$

and

$$\mathbf{Cov}_1(x, x') = \left(1 - \frac{1}{N}\right)\sum_{j=1}^{\infty}\frac{\alpha_j}{j!}\varphi^{(j-1)}(x)\varphi^{(j-1)}(x'), \quad (2)$$

where $\varphi^{(j)}(x)$ denotes the jth derivative of the standard normal density $\varphi(x)$, and

$$\alpha_j = \frac{1}{M}\sum_{i<i'}\rho_{ii'}^j = \int_{-1}^{1}\rho^j\, dG(\rho)$$

is the jth moment according to the empirical measure G of the $M = N(N-1)/2$ pairwise correlations $\rho_{ii'}$.

In Proposition 1, as in Theorem 1 of Efron, $\mathbf{Cov}_0$ is the covariance under independence and $\mathbf{Cov}_1$ is the additional covariance as a result of correlation. The empirical measure G corresponds to the "density" $g(\rho)$ in (2.16) of Efron. Taking a second-order approximation to (2), as Efron suggests in (2.33), Proposition 1 implies particularly that

$$\operatorname{Var}[\hat{F}(x)] \doteq \frac{1}{N}\Phi^+(x)[1 - \Phi^+(x)] + \left(1 - \frac{1}{N}\right)\left(\alpha_1 + \frac{\alpha_2}{2}x^2\right)\varphi^2(x).$$

Notice that α_1 need not be zero as in Efron's approximations if the data does not come from a column-normalized matrix, but in general,

$$-\frac{1}{N-1} \leq \alpha_1 \leq 1, \qquad 0 \leq \alpha_2 \leq 1, \quad (3)$$

where the lower bound on α_1 is a direct consequence of the fact that the matrix of correlations is nonnegative definite (see Appendix). The fact that α_1 is most likely nonnegative for large N implies that the observed histogram will more likely be narrow than wide in most unnormalized situations, as we will see below.

When extended to the multiclass model, the above results are useful for assessing the variability of the empirical distribution $\hat{F}(x)$, leading to expressions such as (1.4) of Efron. I would like to study instead the effect of correlation on the shape that the observed distribution $\hat{F}(x)$ and the corresponding histogram may take. To keep things simple, I will work with the "one-class" null model rather than the multiclass model.

For large N, the covariance function of Proposition 1 approximates the covariance function

$$C(x, x') = \sum_{j=1}^{\infty}\frac{\alpha_j}{j!}\varphi^{(j-1)}(x)\varphi^{(j-1)}(x'). \quad (4)$$

Thus, $\hat{F}(x)$ approximates a random function $F_0(x)$ with mean $\Phi^+(x)$ and covariance function $C(x, x')$. This covariance structure suggests an expansion of $F_0(x)$ in terms of the basis $\varphi^{(0)}(x), -\varphi^{(1)}(x), \varphi^{(2)}(x), \ldots$ as

$$F_0(x) = \Phi^+(x) + \sum_{j=1}^{\infty} W_j(-1)^j\varphi^{(j-1)}(x), \quad (5)$$

where $W_1, W_2, \ldots$ are random variables with mean 0 and covariance

$$\operatorname{cov}(W_j, W_{j'}) = \begin{cases} \alpha_j/j!, & j = j' \\ 0, & j \neq j' \end{cases} \quad (6)$$

obtained by setting $\operatorname{cov}[F_0(x), F_0(x')] = C(x, x')$. Additional constraints exist between the weights $W_1, W_2, \ldots$ in order to ensure that $F_0(x)$ is nondecreasing.

Equation (5) tells us how to generate likely outcomes of the empirical cdf $\hat{F}(x)$ when N is large. A better indication for what these distributions may look like may be obtained by differentiating (5) to get

$$f_0(x) = \varphi(x) + \sum_{j=1}^{\infty} W_j(-1)^j\varphi^{(j)}(x). \quad (7)$$

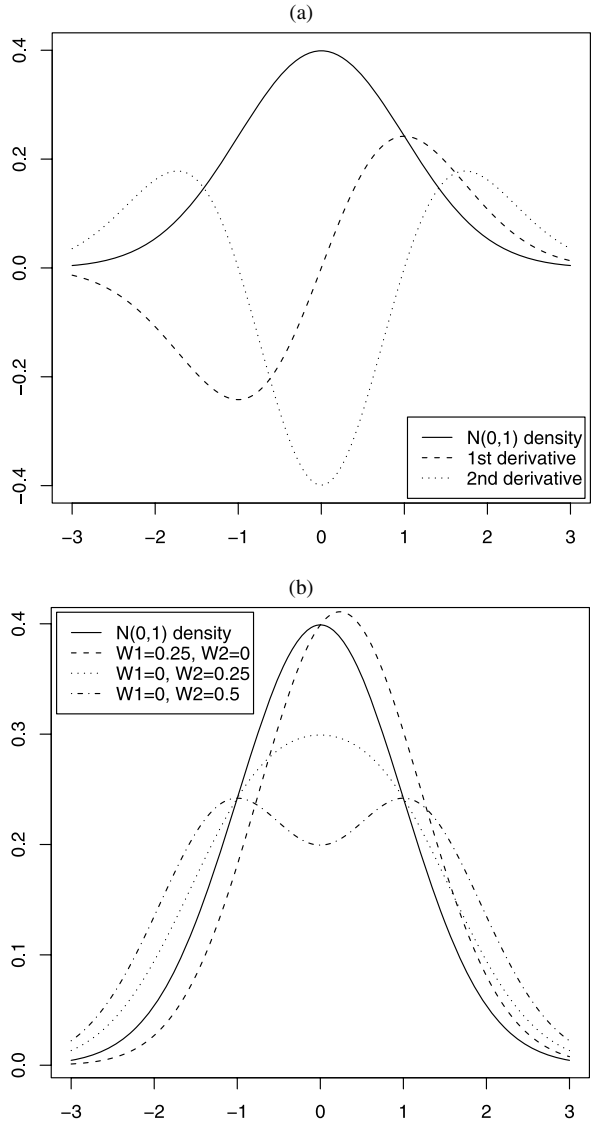

Figure 1. (a) First three basis functions of the Gram–Charlier expansion (7). (b) Examples of realizations of the random density (7).

Equation (7) can be recognized as a Gram–Charlier series expansion of type A (Cramér 1957; Blinnikov and Moessner 1998). It says that the observed N(0, 1) histogram is modified by adding the basis functions $-\varphi^{(1)}(x), \varphi^{(2)}(x), \ldots$, randomly scaled. Notice that these are not quite principal components because the basis functions are not orthogonal in an L_2 sense, but still give a description of the main modes of variation. In the special case of column-normalized data, $\alpha_1 = 0$ forces $W_1 = 0$ by (6), and (7) resembles equation (49) of Efron (2007a).

Figure 1 shows the first three basis functions and a few examples of $f_0(x)$. The first basis function $-\varphi^{(1)}(x)$ produces a shift in the density, to the left if $W_1 < 0$ and to the right if $W_1 > 0$. The second basis function $\varphi^{(2)}(x)$ produces narrowing if $W_2 < 0$ or widening if $W_2 > 0$. It may also produce a bimodal distribution if $W_2 > 0$ is large enough. The higher order basis functions have less of an effect because the variance of $W_3, W_4, \ldots$ decays as $1/j!$ by (6). However, they may be important for high correlation. In the extreme case of perfect correlation where $\rho_{ii'} = 1$ for all i, i', we have that $\alpha_j = 1$ for all j, and $F_0(x)$ takes the shape of a step function where the location of the step has a standard normal distribution. The Gram–Charlier expansion is known to have poor convergence properties (Cramér 1957; Blinnikov and Moessner 1998), and it is not clear whether it would converge in this case.

To answer the question of whether correlation can produce substantial widening of the N(0, 1) density, it is convenient to write (7) as

$$f_0(x) = \varphi(x) \sum_{j=0}^{\infty} W_j h_j(x), \tag{8}$$

where $W_0 = 1$ and $h_j(x)$ denotes the jth Hermite polynomial $h_0(x) = 1$, $h_1(x) = x$, $h_2(x) = x^2 - 1, \ldots$ as in (2.15) of Efron. The derivatives of $f_0(x)$ are given by

$$f_0^{(l)}(x) = (-1)^l \varphi(x) \sum_{j=0}^{\infty} W_j h_{j+l}(x), \qquad l = 0, 1, 2, \ldots. \tag{9}$$

Consider now the moments of the random density $f_0(x)$, themselves random. Using (8) it is easy to show that the weights W_j can be interpreted as the "Hermite moments" of $f_0(x)$,

$$W_j = \frac{1}{j!} \int_{-\infty}^{\infty} h_j(x) f_0(x)\, dx, \qquad j = 0, 1, 2, \ldots. \tag{10}$$

In particular, we can compute the random mean and variance of $f_0(x)$:

$$\begin{aligned} \hat{\mu}_0 &= \int_{-\infty}^{\infty} x f_0(x)\, dx = W_1, \\ \hat{\sigma}_0^2 &= \int_{-\infty}^{\infty} x^2 f_0(x)\, dx - \left[\int_{-\infty}^{\infty} x f_0(x)\, dx \right]^2 \\ &= 2W_2 + 1 - W_1^2. \end{aligned} \tag{11}$$

We see that W_1 controls the mean $\hat{\mu}_0$ of $f_0(x)$ directly. However, the variance $\hat{\sigma}_0^2$ of $f_0(x)$ is controlled by both W_1 and W_2. Its expected value is

$$E(\hat{\sigma}_0^2) = E(2W_2 + 1 - W_1^2) = 1 - \alpha_1$$

by (6). If $\alpha_1 = 0$, as forced by the column normalization, then $E(\hat{\sigma}_0^2) = 1$, meaning that the observed histogram can be sometimes wider or sometimes narrower than N(0, 1). On the other hand, if $\alpha_1 > 0$, then $E(\hat{\sigma}_0^2) < 1$ and we would expect a *narrower* histogram than N(0, 1). In order to expect a wider histogram, we would need $\alpha_1 < 0$, but this is unlikely for large N by (3).

As shown in Figure 1, if $f_0(x)$ is indeed wider than N(0, 1), it cannot be very wide before it becomes bimodal. As an approximation, discarding terms for $j \geq 4$ in (9), we get that the curvature of $f_0(x)$ at zero,

$$\begin{aligned} f^{(2)}(0) &\doteq \varphi(0)[h_2(0) + W_2 h_4(0)] \\ &= \varphi(0)[-1 + 3W_2], \end{aligned}$$

remains nonpositive (concave) only for $W_2 \leq 1/3$. Plugging in (11), we get that the largest central spread for a unimodal density is about $\hat{\sigma}_0 \approx 1.3$.

This result is important because it implies that the central spread $\hat{\sigma}_0 = 1.68$ of the observed histogram for the leukemia data in Figure 1 of Efron's current paper cannot be explained

on the basis of correlation alone. The same is true for the central spread $\hat{\sigma}_0 = 1.55$ of the observed histogram for the breast cancer data in figure 1 of Efron (2007a). The sampling variability in the estimation of the true correlations $\rho_{ii'}$ may affect the estimation of the correlation parameters α_j, but does not affect the spread of the histogram of z variates, which is specified by the random variables W_j.

A possible explanation for the unusually large central spread in could be the presence of unobserved covariates, as suggested in Efron (2007b). Otherwise, it may be that the nonnull component of the mixture is stronger than shown and that the nonnull counts (the red bars in Figure 1 of Efron's current paper) should extend further into the the center of the histogram. This is impossible to detect if the empirical null is purposefully made to match the center of the histogram. Assuming a null distribution that is wider than it should be will result in a loss of detection power.

3. THE DISTRIBUTION OF CORRELATED χ^2 VARIATES

As an additional issue, I would like to comment on Efron's recommendation to transform nonnormal variates to a normal scale. In Section 5 of the current paper, Efron uses the central limit theorem to argue that both null and nonnull nonnormal variates can look remarkably normal when transformed to a normal scale via a quantile transformation such as (5.1) or (5.22). While this is true in the t and gamma examples presented, there are situations where the argument does not apply. For example, in likelihood ratio tests and the second-order delta method, the asymptotic null distribution of the test statistics for large sample size is χ^2 rather than normal.

The question is whether a transformation like (5.22) should be applied and then the analysis performed on the z variables. Working with normal variates is easier and allows taking advantage of that theory. On the other hand, the quantile transformation (5.22) might substantially shift the mode for the purposes of fitting an empirical null (Schwartzman 2008). In addition, it is unclear how the correlation structure is mapped and whether the resulting normal variates would still be jointly normal.

An idea suggested in Schwartzman and Lin (2009) is to model the χ^2 test statistics directly. Let $f_\nu(t)$ denote the density of the $\chi^2(\nu)$ distribution with ν df. Under the complete null, the pair of variables $(T_i, T_{i'})$ admits a Lancaster bivariate model where both T_i and $T_{i'}$ have the same marginal density f_ν, their correlation is $\rho_{ii'}$, and their joint density is given by

$$f_\nu(t_i, t_{i'}; \rho_{ii'}) = f_\nu(t_i) f_\nu(t_{i'}) \sum_{k=0}^{\infty} \frac{\rho_{ii'}^k}{k!} \frac{\Gamma(\nu/2)}{\Gamma(\nu/2+k)} \times \mathcal{L}_k^{(\nu/2-1)}\left(\frac{t_i}{2}\right) \mathcal{L}_k^{(\nu/2-1)}\left(\frac{t_{i'}}{2}\right), \quad (1)$$

where $\mathcal{L}_k^{(\nu/2-1)}(t)$ are the generalized Laguerre polynomials of degree $\nu/2-1$: $\mathcal{L}_0^{(\nu/2-1)}(t) = 1$, $\mathcal{L}_1^{(\nu/2-1)}(t) = -t + \nu/2$, $\mathcal{L}_2^{(\nu/2-1)}(t) = t^2 - 2(\nu/2+1)t + (\nu/2)(\nu/2+1)$ and so on (Koudou 1998). A derivation similar to the one above for the Gaussian case gives that, for large N, the covariance function (4) takes the form

$$C(x, x') = f_{\nu+2}(x) f_{\nu+2}(x') \sum_{j=1}^{\infty} \frac{\alpha_j}{j!} \frac{\Gamma(\nu/2)}{\Gamma(\nu/2+k)} \times \mathcal{L}_{k-1}^{(\nu/2-1)}\left(\frac{x}{2}\right) \mathcal{L}_{k-1}^{(\nu/2-1)}\left(\frac{x'}{2}\right) \quad (2)$$

and the expansions (7) and (8) become

$$f_0(x) = f_\nu(x) + \sum_{j=1}^{\infty} W_j \sqrt{\frac{\Gamma(\nu/2+k)}{\Gamma(\nu/2)}} f_{\nu+2k}^{(j)}(x) = f_\nu(x) \sum_{j=0}^{\infty} W_j \sqrt{\frac{\Gamma(\nu/2)}{\Gamma(\nu/2+k)}} \mathcal{L}_k^{(\nu/2-1)}\left(\frac{x}{2}\right), \quad (3)$$

where $f_\nu^{(j)}(x)$ denotes the jth derivative of the density $f_\nu(x)$, $W_0 = 1$ and $W_1, W_2, \ldots$ are uncorrelated random variables with mean 0 and variance $\alpha_j/j!$.

4. SUMMARY

In this comment, I have attempted to exploit Efron's results to shed some light on the effect that correlation has on the observed distribution of a large collection of variables, first standard normal and then χ^2. To do this, I derived a simplified "complete null" version of Efron's Theorem 1 working directly with empirical distribution functions rather than histograms, and then suggested an approximation to the observed density as a Gram–Charlier expansion in terms of Hermite polynomials in the normal case and generalized Laguerre polynomials in the χ^2 case. Beter approximations may exist based on Gauss–Hermite or Edgeworth expansions (Cramér 1957; Blinnikov and Moessner 1998).

One of the conclusions is that correlation modifies the common marginal density by adding or subtracting the density's derivatives multiplied by random coefficients. These coefficients have zero mean, are dependendent but uncorrelated, and their variance is proportional to the empirical moments of the pairwise correlations, decaying as the order increases. This representation suggests that correlation is more likely to narrow the observed histogram of test statistics than widen it, and that the observed histogram cannot be too wide before it becomes bimodal.

While I have only considered the "one-class" model, the results extend to Efron's normal multiclass model under the assumption that the "null" class dominates the mixture and that the other classes have little overlap with the mode. This assumption is the same one made in Efron (2007a) in order to fit an empirical null. More generally, a multiclass version of (4) for normal variables would resemble the first term of Efron's (2.31) and a corresponding version of (7) could be proposed. Unfortunately, this could not be done easily for χ^2 variables as model (1) cannot be easily extended to the noncentral case.

I want to conclude remarking that whether an empirical null should be used at all is still an open question. In Section 4 of the current article, Efron states that estimating and empirical null "is both necessary and feasible but can greatly increase variability." The feasibility and increased variability are understood. The necessity, however, is a more fundamental question

of marginal versus conditional inference: marginal over all possible realizations of the data or conditional on the particular data observed. This is a crucial question for large-scale hypothesis testing problems. Who better than Efron can answer such questions? I look forward to further developments in this area.

APPENDIX: PROOFS AND DERIVATIONS

Proof of Proposition 1

For two standard normal variables Z, Z' with correlation ρ, let $\Phi^+_\rho(x,x') = P(Z \geq x, Z' \geq x')$ be the corresponding right-tail bivariate cdf. Then

$$\begin{aligned}\operatorname{cov}[\hat{F}(x), \hat{F}(x')] &= E[\hat{F}(x)\hat{F}(x')] - E[\hat{F}(x)]E[\hat{F}(x')] \\ &= \frac{1}{N^2}\sum_{i,i'} P(Z_i \geq x, Z_{i'} \geq x') - \Phi^+(x)\Phi^+(x') \\ &= \frac{1}{N^2} N\Phi^+(\max(x,x')) \\ &\quad + \frac{1}{N^2}\sum_{i\neq i'} \Phi^+_{\rho_{ii'}}(x,x') - \Phi^+(x)\Phi^+(x') \\ &= \mathbf{Cov}_0(x,x') + \mathbf{Cov}_1(x,x'),\end{aligned}$$

where $\mathbf{Cov}_0(x,x')$ is given by (1) and

$$\mathbf{Cov}_1(x,x') = \frac{1}{N^2}\sum_{i\neq i'} \Phi^+_{\rho_{ii'}}(x,x') - \left(1 - \frac{1}{N}\right)\Phi^+(x)\Phi^+(x'). \quad \text{(A.1)}$$

To obtain (2), we proceed as follows. As in (2.10) of Efron, let $\varphi_\rho(x,x')$ denote the bivariate standard normal density with correlation ρ. Mehler's identity (Patel and Read 1996; Kotz, Balakrishnan, and Johnson 2000), also used in (2.15) of Efron, states that

$$\begin{aligned}\varphi_\rho(x,x') &= \varphi(x)\varphi(x')\sum_{j=0}^{\infty}\frac{\rho^j}{j!}h_j(x)h_j(x') \\ &= \varphi(x)\varphi(x') + \sum_{j=1}^{\infty}\frac{\rho^j}{j!}\varphi^{(j)}(x)\varphi^{(j)}(x'),\end{aligned}$$

where $h_j(x)$ is the jth Hermite polynomial and the jth derivative satisfies $\varphi^{(j)}(x) = (-1)^j\varphi(x)h_j(x)$. Integrating from x to ∞ and x' to ∞, we get that

$$\Phi^+_\rho(x,x') = \Phi^+(x)\Phi^+(x') + \sum_{j=1}^{\infty}\frac{\rho^j}{j!}\varphi^{(j-1)}(x)\varphi^{(j-1)}(x').$$

Replacing in (A.1) gives

$$\begin{aligned}\mathbf{Cov}_1(x,x') &= \frac{1}{N^2}\sum_{i\neq i'}\left[\Phi^+(x)\Phi^+(x') + \sum_{j=1}^{\infty}\frac{\rho_{ii'}^j}{j!}\varphi^{(j-1)}(x)\varphi^{(j-1)}(x')\right] \\ &\quad - \left(1 - \frac{1}{N}\right)\Phi^+(x)\Phi^+(x') \\ &= \frac{1}{N^2}\sum_{j=1}^{\infty}\frac{1}{j!}\left(\sum_{i\neq i'}\rho_{ii'}^j\right)\varphi^{(j-1)}(x)\varphi^{(j-1)}(x'),\end{aligned}$$

yielding (2).

Derivation of (3)

The upper bounds are immediate from the definition of α_1 and α_2. To see the lower bound on α_1, let $\mathbf{R} = \{\rho_{ii'}\}$ be the correlation matrix of the vector $(Z_1, \ldots, Z_N)^T$ and let $\mathbf{1}$ be a column vector of ones. Since $\mathbf{1}^T\mathbf{R}\mathbf{1} \geq 0$ by the positive semidefiniteness of $\mathbf{R}$,

$$\alpha_1 = \frac{1}{N(N-1)}\sum_{i\neq i'}\rho_{ii'} = \frac{1}{N(N-1)}(\mathbf{1}^T\mathbf{R}\mathbf{1} - N) \geq -\frac{1}{N-1}.$$

ADDITIONAL REFERENCES

Blinnikov, S., and Moessner, R. (1998), "Expansions for Nearly Gaussian Distributions," *Astronomy and Astrophysics Supplement Series*, 130, 193–205. [1061,1062]

Cramér, H. (1957), *Mathematical Methods of Statistics*, Princeton: Princeton University Press. [1061,1062]

Efron, B. (2004), "Large-Scale Simultaneous Hypothesis Testing: The Choice of a Null Hypothesis," *Journal of the American Statistical Association*, 99, 96–104. [1059]

——— (2008), "Simultaneous Inference: When Should Hypothesis Testing Problems Be Combined?" *The Annals of Applied Statistics*, 2, 197–223. [1059]

Kotz, S., Balakrishnan, N., and Johnson, N. L. (2000), "Bivariate and Trivariate Normal Distributions," in *Continuous Multivariate Distributions, Volume 1: Models and Applications*, New York: Wiley, pp. 251–348. [1063]

Koudou, A. E. (1998), "Lancaster Bivariate Probability Distributions With Poisson, Negative Binomial and Gamma Margins," *Test*, 7, 95–110. [1062]

Patel, J. K., and Read, C. B. (1996), *Handbook of the Normal Distribution* (2nd ed.), New York: Dekker. [1063]

Schwartzman, A. (2008), "Empirical Null and False Discovery Rate Inference for Exponential Families," *The Annals of Applied Statistics*, 2, 1332–1359. [1062]

Comment

Peter H. WESTFALL

1. INTRODUCTION

The scientific literature has recently experienced an embarrassment of contradictory results (Ioannidis 2005; Bertram et al. 2007; Boffetta 2008). Such findings are costly to the statistical profession that provides methods, and they are costly to the scientific enterprise in general. Validity of inference is therefore important. Professor Efron has once again done the statistical community a great service by quantifying the accuracy of the estimates of complex inferential quantities, and it is a privilege to comment on his article.

While the main application of the paper is to multiple comparisons, the actual scope is, as the title suggests, much broader.

Peter H. Westfall is Horn Professor of Statistics, Area of ISQS, Texas Tech University, Lubbock, TX 79409 (E-mail: *peter.westfall@ttu.edu*).

Journal of the American Statistical Association
September 2010, Vol. 105, No. 491, Theory and Methods
DOI: 10.1198/jasa.2010.tm10239

The paper has two main theoretical results, one about the accuracy of estimates of distribution functions related to measures of evidence in multiple comparisons, and the other about asymptotic nonnull distributions of z statistics. The two results stand independently of each other as useful general contributions, but together tell a coherent tale, with unifying application to gene expression studies. While the paper has wide-ranging applications, I will confine my comments to multiple comparisons.

Degradation of inference caused by deviation from independently and identically distributed (iid) *observations* has been studied extensively in the literature, mostly in the case of standard (not multiple) inferences. Efron's article considers how and whether similar degradation occurs when *variables* are correlated, when there are multiple inferences, one per variable. The question is relevant since much of the multiple comparisons literature assumes independence, and some literature suggests that multiple comparisons procedures derived under independence are reasonably robust to violations of independence. A main conclusion of Efron's paper is that, on the contrary, correlations do matter.

The conclusion that correlations are important in multiple comparisons procedures is no surprise; this is the main reason that Tukey's and Dunnett's classical methods are used instead of the Bonferroni method for familywise error rate (FWER) control. More recent literature involves using correlations to improve power in more complex applications, for example by resampling vectors in nonparametric settings (Westfall and Young 1993), or by using quasi-Monte Carlo methods in parametric settings (Genz and Bretz 2009). It is well known that the Bonferroni method is nearly correct under independence since $\alpha/N \simeq 1 - (1-\alpha)^{1/N}$, but that critical p-value thresholds for scan-type statistics can be reduced well below α/N to, say, α/N^*, $1 \le N^* \le N$, where $N^* \ll N$ under strong dependence. It is also known that N^* is not a linear function of "size" of correlation: for typical correlation structures N^* is typically not far from N; substantial reductions occur only for extreme dependencies such as with multiple comparisons of response surfaces (e.g., Liu, Jamshidian, and Zhang 2004). Similar reductions in conservativeness FDR-controlling procedures when correlation information is incorporated have also been noted (e.g., Troendle 2000).

Professor Efron's paper implicitly has the *opposite* conclusion, suggesting that multiple comparisons procedures should be *more* conservative with larger correlations. This conclusion follows from the correlation penalty that is the second term in (1.4), and its effect on assessed accuracy of statistics that are used to assess whether a test (gene, e.g.) should be flagged as "interesting." With greater uncertainty in the estimate of the measure of "interestingness," the natural response would be to adopt a more conservative stance, perhaps taking the upper limit of the confidence interval as the measure of evidence of "interestingness" (smaller values being more interesting), rather than the estimate itself. This approach is implicit in the discussion of the upper limit of the estimate of Fdr(5) for the leukemia data following Equation (4.5). The logical extension: since the upper limit increases with larger correlations, one should be more cautious in flagging tests when there is greater correlation.

With such diametrically opposed conclusions regarding the effect of correlations on multiple inference, some reconciliation is needed.

2. PERMUTATIONS

First, a comment about permutation methods. The comment "[permutation methods] are inappropriate for the accuracy considerations of this paper," as well as the inclusion of the permutation standard errors in Table 1, seems possibly misleading. The permutation standard error in Table 1 involves an artificially created null reference distribution that seems inappropriate for the problem at hand. A better comparator would be the standard nonparametric bootstrap, although Efron notes in Remark F that the bootstrap produces erratic results.

The comment about inappropriateness of permutation methods also seems possibly misleading to the reader who is not fully aware of the context. Large scale correlation structures are incorporated with finite-sample *exact* tests, not asymptotic approximations, to control the familywise error rate (FWER) using permutation methods under a mild exchangeability assumption (Westfall and Troendle 2008). The exactness of these tests occurs, interestingly, despite the gross singularity of the $N \times N$ correlation matrix estimated from n observations when $N \gg n$. Conditional on the appropriate permutation orbit, there is no error at all in the multiplicity-adjusted p-values, assuming that one can enumerate all permutations; otherwise there is Monte Carlo that can be made arbitrarily small. The required exchangeability assumption needed for exactness of the permutation tests is related to the assumption that the correlation distribution $g(\rho)$ is constant across groups (made in Remark C): both assumptions are valid in the multivariate location shift model.

Approximate methods based on the nonparametric bootstrap also incorporate correlation structures, without needing the exchangeability assumption, but behave more erratically (Troendle, Korn, and McShane 2004). No matter whether multivariate bootstrap or multivariate permutation methods are used, however, the net effect of greater correlation is always to reduce conservativeness: the greater the correlation, the greater the reduction.

3. EFFECT OF CORRELATION ON MINIMIZING LOSS

With the "-omics" revolution, there has been an explosion of literature in multiple comparisons procedures. Responding to concerns of conservatism of FWER-controlling procedures with large N, many of these methods involve variants of FDR-controlling methods, with names pFDR, FDR-k, FDP, adaptive FDR, empirical null FDR, local FDR, and so on. The variety of methods can be bewildering to the scientist, who asks the simple question "which method is best?" The answer involves relative severity of Type I to Type II errors. While FWER-controlling methods have become less popular, Baker and Kramer (2008) argue that Type I errors are quite costly, and suggest using FWER-controlling methods. On the other hand, for exploratory applications where Type I errors have little cost, FDR-controlling methods may be more appropriate.

Decision theory provides a framework within which the relative losses of Type I and Type II errors can be incorporated directly. Different loss functions favor different methods; even the Bonferroni method is optimal when a single Type I error is as costly as βN Type II errors (Lu and Westfall 2009). Identifying reasonable loss functions is nontrivial, and requires much discussion and consensus among scientists and statisticians. Even

if loss functions are not formally elicited for specific data analyses, it is still useful to know which multiple comparisons methods are preferred for which types of loss functions, to help decide which method to use.

3.1 A Model for Random Correlation

What follows is a highly simplified model of a gene expression study, one that is perhaps more realistic for studying the effects of correlation than the "two-point correlation distribution" simulation model [Equation (2.24); see also Efron's Remark D], that can shed light on the effects of correlation on loss. This model has several useful properties: (i) it generates correlations with mean zero, (ii) it is simple to generate data with any desired root mean square correlation, (iii) it generates correlations in the continuum $(-1, 1)$, and (iv) it guarantees positive definiteness of the correlation matrix.

Suppose the two-class model as in (2.4)–(2.5), $Z_i \sim \mathrm{N}(\mu_i, 1)$ and that the joint distribution over all genes is $\mathbf{Z} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Phi})$. Assume $\boldsymbol{\Phi} = \boldsymbol{\lambda}\boldsymbol{\lambda}' + \boldsymbol{\Psi}^2$, where $\boldsymbol{\lambda}$ is $N \times 1$ and $\boldsymbol{\Psi}^2 = \mathrm{diag}\{\psi_i^2\}$. Then

$$\rho_{ij} = \mathrm{corr}(Z_i, Z_j) = \lambda_i \lambda_j. \tag{1}$$

We wish to view these correlations as random, with mean zero, and with a second moment that can be specified. Inject randomness as follows: define $U_i \sim_{\mathrm{iid}} U(-1, 1)$, and let $\lambda_i = U_i/(U_i^2 + s^2)^{1/2}$; s^2 is a parameter to be chosen. Then $E(\rho_{ij}) = 0$ and $\mathrm{rmsc} \equiv \{E(\rho_{ij}^2)\}^{1/2} = 1 - s\tan^{-1}(1/s)$; smaller s imply greater dependence. Figure 1 displays the relationship between s and rmsc.

3.2 Loss Functions

The question to be addressed is, "should we be more conservative when there is more correlation?" Consider then the effect of dependence structure on optimal decision making. The decisions are to either to declare that transcript i is "underexpressed" (UE), "overexpressed" (OE), or "not interesting" (NI). These decisions are made when $Z_i < -c_i$, $Z_i > c_i$, or $|Z_i| \le c_i$, respectively; the goal is to choose the "best" c_i. Loss associated with classifying transcript i using threshold c_i is

$$L(Z_i, c_i) = I(Z_i < -c_i)L_{\mathrm{UE}}(\mu_i) + I(Z_i > c_i)L_{\mathrm{OE}}(\mu_i) + I(|Z_i| \le c_i)L_{\mathrm{NI}}(\mu_i)$$

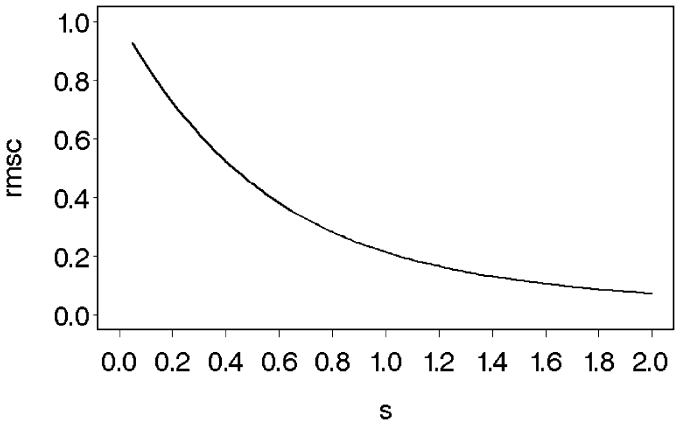

Figure 1. Root mean square correlation (rmsc) as a function of s in the simulation model (1) with randomly generated λ.

for loss functions $L_{\mathrm{UE}}(\mu)$, $L_{\mathrm{OE}}(\mu)$ and $L_{\mathrm{NI}}(\mu)$. The Waller–Duncan (WD) loss functions are popular:

$$L_{\mathrm{UE}}^{\mathrm{WD}}(\mu) = K\mu I(\mu \ge 0),$$
$$L_{\mathrm{OE}}^{\mathrm{WD}}(\mu) = -K\mu I(\mu \le 0), \qquad \text{and}$$
$$L_{\mathrm{NI}}^{\mathrm{WD}}(\mu) = |\mu|,$$

where K denotes the severity of a Type I error relative to a Type II error. Because the term μ appears in the Type I error losses $L_{\mathrm{UE}}(\mu)$ and $L_{\mathrm{OE}}(\mu)$, there is essentially zero Type I error penalty in the case of the two-class model (2.24) where $\mu = 0$ for 95% of the cases. In this case one might remove the μ terms from all loss functions:

$$L_{\mathrm{UE}}^{\mathrm{NM}}(\mu) = KI(\mu \ge 0),$$
$$L_{\mathrm{OE}}^{\mathrm{NM}}(\mu) = KI(\mu \le 0), \qquad \text{and}$$
$$L_{\mathrm{NI}}^{\mathrm{NM}}(\mu) = I(\mu \ne 0),$$

with "NM" denoting "no mu."

There are several ways to combine loss from individual actions arrive at an "overall" loss; additive loss is common:

$$L(\mathbf{Z}, \mathbf{c}) = \sum_i \big\{ I(Z_i < -c_i)L_{\mathrm{UE}}(\mu_i) + I(Z_i > c_i)L_{\mathrm{OE}}(\mu_i) + I(|Z_i| \le c_i)L_{\mathrm{NI}}(\mu_i) \big\}. \tag{2}$$

As $L(\mathbf{Z}, \mathbf{c})$ is random, a property of its distribution must be selected first before choosing the optimal c_i. Typically the expected value is minimized; however, minimizing $E\{L(\mathbf{Z}, \mathbf{c})\}$ can be accomplished by minimizing the component expectations $E\{L(Z_i, c_i)\}$ (e.g., Pennello 1997). So correlation simply doesn't matter in this case. If correlation really matters, then the decision maker must have either another property of the loss distribution in mind or a different loss function, for which expected loss depends on correlation.

In financial analysis, "Value at Risk" (abbreviated VaR, Jorion 2007) is used as an upper bound on the loss anticipated from an investment portfolio. The 0.95 and 0.99 quantiles are typical. It is well known that the correlations among portfolio items contribute to VaR: higher correlations imply higher VaR, providing a fundamental insight about diversification. The same applies here as well: dependence among Z_i affects the variance of the sum in (2), which in turn affects the quantiles. So define the VaR-like quantiles $q^{0.95}\{L(\mathbf{Z}, \mathbf{c})\}$ and $q^{0.99}\{L(\mathbf{Z}, \mathbf{c})\}$.

Alternatively, one may consider loss functions other than additive for which expected loss does depend on correlation; this might be a fruitful avenue for further research.

4. SIMULATION STUDY

Datasets $\{\mathbf{Z}\}$ were simulated using the model described in the previous section, with the intent of mimicking the simulation described in (2.24) of Efron, as follows:

1. Pick a desired rmsc, solve for s.
2. Simulate (once, prior to main simulation loop) $\lambda_i = U_i/(U_i^2 + s^2)^{1/2}$, $i = 1, \ldots, 6000$, where $U_i \sim_{\mathrm{iid}} U(-1, 1)$.
3. Simulate (once, prior to main simulation loop) $\mu_i \sim_{\mathrm{iid}}$, $i = 1, \ldots, 6000$, from the two-point distribution with weights $(0.95, 0.05)$ on $(0, 2.5)$.

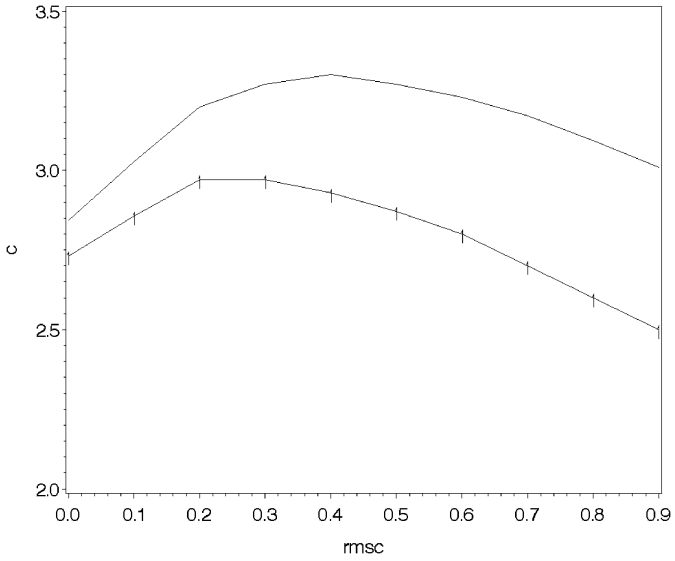

Figure 2. Estimated critical values c that minimize 0.95 quantile (hatched) and 0.99 quantile (smooth) of L^{NM} loss function with $K = 1$, as a function of root mean square correlation (rmsc). Plots are LOESS smoothed.

4. Generate $F \sim \mathrm{N}(0, 1)$ and $\varepsilon_1, \ldots, \varepsilon_{6000} \sim_{\mathrm{iid}} \mathrm{N}(0, 1)$, independent of F. Set $Z_i = \mu_i + \lambda_i F + (1 - \lambda_i^2)^{1/2}\varepsilon_i$, $i = 1, \ldots, 6000$.
5. Find the loss (2) using $c_i \equiv c$, with $c = 1.5, 1.6, \ldots, 4.0$. (Note that common random numbers are used to minimize variation in Monte Carlo estimates of loss for the various c.)
6. Repeat 4–5 5000 times, and identify the c that minimize the estimates of $q^{0.95}\{L(\mathbf{Z}, c)\}$ and $q^{0.99}\{L(\mathbf{Z}, c)\}$ based on these 5000 simulations.

Figure 2 shows the estimated c that minimize $q^{0.95}\{L^{\mathrm{NM}}(\mathbf{Z}, c)\}$ and $q^{0.99}\{L^{\mathrm{NM}}(\mathbf{Z}, c)\}$ ($K = 1$) as a function of rmsc. The c that minimizes $E\{L^{\mathrm{NM}}(\mathbf{Z}, c)\}$ was estimated at a constant $c = 2.7$, independent of rmsc as expected, and is not shown.

Similar results are obtained for other K in L^{NM}. The simulation also was modified so that the L^{WD} loss function makes sense: in step 3, the μ_i were generated as $0.95\mathrm{N}(0, 1) + 0.05\mathrm{N}(2.5, 1)$, and similar behavior as in Figure 2 was noted. In the case $K = 500$ with L^{WD}, the c decreased monotonically in rmsc, suggesting that one should be less conservative with any increase in correlation.

Based on the simulations, I conclude that I should not necessarily be more conservative with larger correlations, no matter whether expected or upper quantile of loss is minimized. So, while I believe that the accuracy of the estimates of Fdr(z) and fdr(z) depend on correlation, with less accuracy for larger correlation, I am confused as to what I should do with this information.

5. CONCLUSION

A method that does not work well across different data-specific and philosophical modalities is considered less desirable than one that works well across various modalities. The initial impetus for the development of FDR-controlling methods was that FWER-controlling methods do not "scale up" well with increasing N. Subsequently, it was found that statistics like Fdr(z) have Bayesian as well as frequentist interpretations, further increasing their desirability.

The present paper makes clear the fact that the estimates of Fdr(z) and fdr(z) do not "scale down" well. Certainly, as Professor Efron's paper makes clear, larger correlation structure decreases the accuracy of these estimates; but even under independence, the estimates will be inaccurate for small N. On the other hand, there are readily available methods for multiple comparisons that work very well with small N *and* large correlations, and that have nothing to do with estimation of Fdr(z), fdr(z); these methods are routinely required by the U.S. Food and Drug Administration for the evaluation of pharmaceutical products, for example, in the analysis of multiple endpoints.

The more I think about the complex questions in multiple comparisons, the more I think that the best answers are couched in decision theory. But there is much more work to do.

ADDITIONAL REFERENCES

Baker, S. G., and Kramer, B. S. (2008), "Using Microarrays to Study the Microenvironment in Tumor Biology: The Crucial Role of Statistics," *Seminars in Cancer Biology*, 18, 305–310. [1064]

Bertram, L., McQueen, M. B., Mullin, K., Blacker, D., and Tanzi, R. E. (2007), "Systematic Meta-Analyses of Alzheimer Disease Genetic Association Studies: The AlzGene Database," *Nature Genetics*, 39, 17–23. [1063]

Boffetta, P., McLaughlin, J. K., La Vecchia, C., Tarone, R. E., Lipworth, L., and Blot, W. J. (2008), "False-Positive Results in Cancer Epidemiology: A Plea for Epistemological Modesty," *Journal of the National Cancer Institute*, 100, 988–995. [1063]

Genz, A., and Bretz, F. (2009), *Computation of Multivariate Normal and t Probabilities. Lecture Notes in Statistics*, Vol. 195, Heidelberg: Springer. [1064]

Ioannidis, J. P. (2005), "Contradicted and Initially Stronger Effects in Highly Cited Clinical Research," *Journal of the American Medical Association*, 294, 218–228. [1063]

Jorion, P. (2007), *Value at Risk—The New Benchmark for Managing Financial Risk* (3rd ed.), Boston: Springer. [1065]

Liu, W., Jamshidian, M., and Zhang, Y. (2004), "Multiple Comparison of Several Linear Regression Models," *Journal of the American Statistical Association*, 99, 395–403. [1064]

Lu, Y., and Westfall, P. (2009), "Is Bonferroni Admissible for Large m?" *American Journal of Mathematical and Management Sciences*, 29, 51–69. [1064]

Pennello, G. (1997), "The k-Ratio Multiple Comparisons Bayes Rule for the Balanced Two-Way Design," *Journal of the American Statistical Association*, 92, 675–684. [1065]

Troendle, J. F. (2000), "Stepwise Normal Theory Multiple Test Procedures Controlling the False Discovery Rate," *Journal of Statistical Planning and Inference*, 84, 139–158. [1064]

Troendle, J. F., Korn, E. L., and McShane, L. M. (2004), "An Example of Slow Convergence of the Bootstrap in High Dimensions," *The American Statistician*, 58, 25–29. [1064]

Westfall, P. H., and Troendle, J. F. (2008), "Multiple Testing With Minimal Assumptions," *Biometrical Journal*, 50, 745–755. [1064]

Rejoinder

Bradley EFRON

This paper is actually two papers, as Peter Westfall and Armin Schwartzman point out. The first concerns the accuracy of summary statistics based on correlated normal variates, while the second shows that the accuracy theory applies to z-values. Both parts were necessary for my original purpose: to assess the accuracy of $\widehat{\text{fdr}}(z)$ compared to $\widehat{\text{Fdr}}(z)$, that is, of local versus tail area false discovery rate estimates. Figure 6 shows that my first guess, that $\widehat{\text{fdr}}(z)$ would be more variable since it involves density estimation, was wrong: in fact, $\widehat{\text{fdr}}(z)$ is slightly more stable than $\widehat{\text{Fdr}}(z)$. The real story is the greatly increased variability of *both* estimates when correlation enters the picture.

All four discussants, Ruth Heller and Tony Cai as well as Schwartzman and Westfall, have written authoritatively on large-scale testing. The four essays are interesting in their own right, often going beyond the confines of my paper into more general questions of large-scale inference. I cannot deal here with all the points they raise, but will try to comment on some of the more provocative ones.

First though, I wanted to say a few words about a missing topic in the paper: the use of ordinary nonparametric bootstrapping to answer accuracy questions like those of Table 1. The 7128×72 leukemia data matrix $\mathbf{X}$ has 47 ALL and 25 AML columns. Choosing 47 ALL columns randomly and with replacement, and likewise 25 AML columns, gives bootstrap matrix $\mathbf{X}^*$, bootstrap z-values z_i^* for $i = 1, 2, \ldots, 7128$, as for the actual data, and bootstrap cdf $\hat{F}(x)^* = \#\{z_i^* \le x\}/7128$. Repeating this process B times produces bootstrap standard deviation estimates in the usual way,

$$\widehat{\text{sd}}_{\text{boot}}(x) = \left[\sum_{b=1}^{B} (\hat{F}(x)^{*b} - \hat{F}(x)^{*\cdot})^2 \Big/ (B-1)\right]^{1/2}$$

with $\hat{F}(x)^{*\cdot}$ indicating the average. Since columns of $\mathbf{X}$ are resampled intact, correlations among the genes are correctly accounted for.

All of this is a lot simpler, and less parametric, than formula (1.4). The trouble is, it doesn't give very good answers. Figure 8 below shows $\widehat{\text{sd}}_{\text{boot}}(x)$ to be a somewhat dilated version of formula (1.4). (Simulations like those in Figure 3 confirm the dilation effect.)

It is not hard to see why. Each z_i^* is roughly normally distributed around its original value z_i, say $z_i^* \dot{\sim} \mathcal{N}(z_i, \sigma_i^2)$. This gives the bootstrap histogram an extra component of variance,

$$E_*\left\{\sum_1^N z_i^{*2} \Big/ N\right\} = \left\{\sum_1^N z_i^2 \Big/ N\right\} + \left\{\sum_1^N \sigma_i^2 \Big/ N\right\},$$

producing the *bootstrap dilation* effect. I have been unable to contrive a general nonparametric correction for bootstrap dilation.

Permutations do not suffer from dilation effects, but they are not (and never were intended to be) the answer to questions like those posed in Table 1. Used properly, as in Westfall's path-breaking book with Young, permutation algorithms can efficiently account for some, but not all, correlation effects in multiple testing. A rough guide is that permutation methods can estimate correlations but not standard deviations. Something has widened the center of the leukemia z-value histogram in Figure 1, but we won't find out what from permutation calculations, which always lead back to the theoretical $\mathcal{N}(0, 1)$ null distribution.

Correlations between z-values are the bad guys in my paper, degrading the estimation accuracy of summary statistics such as $\hat{F}(x)$. Degraded inference properties ensue, as illustrated in Heller's figure 1. (The current paper is more probabilistic than statistical. Its predecessor, Efron 2007a, explores the kinds of inferential effects seen in Heller's simulation experiment.)

Cai, Heller, and Westfall raise the possibility of "good" correlation effects, those that improve inferences. Suppose, as Heller suggests, the z-values are related spatially, say with z_i observed at location x_i, and that $z_i \sim \mathcal{N}(\mu_i, \sigma^2)$, where the unobserved μ_i's vary smoothly with location. (This is the setup in Benjamini and Heller 2007.) Nearby z_i's will now tend to appear positively correlated. This encourages averaging the z-values over local neighborhoods in order to improve μ_i's estimation. Cai makes the same suggestion for clustered genes in a microarray study.

All of this would reduce to an ordinary regression problem if μ were known to be a simple function of x, say $\mu_i = \beta_0 + \beta_1 x_i$. Correlations can play good-guy roles in suggesting regression-like structures. Presumably, the kinds of correlations my paper is worrying about are those remaining after the statistician has accounted for such structures. In this context I suspect, but cannot show, that increased correlation always reduces the amount of information available to the statistician. Perhaps Westfall's decision theoretic approach will prove me wrong.

I agree with Schwartzman that correlation by itself is not to blame for all the overdispersion seen in Figure 1. Correlation's effect on $\hat{\sigma}_0$ is limited to a factor of about $1 \pm 2\alpha$ [equations (38) and (60) in Efron 2007a], roughly in the range (0.8, 1.2) for the leukemia data. Unobserved covariates, whose effects would be removed by regression if the covariates were known, are likely suspects.

I disagree, however, that correlation effects on $\hat{\sigma}_0$ are negligible. Figure 9 shows histograms of $\hat{\sigma}_0$ for vector $\mathbf{z}$ having $z_i \sim \mathcal{N}(0, 1)$ for $i = 1, 2, \ldots, 1000$, with root mean square correlation $\alpha = 0.10$ and also for $\alpha = 0.20$. Values of $\hat{\sigma}_0$ *less* than

Bradley Efron is Professor, Department of Statistics, Stanford University, Stanford, CA 94305 (E-mail: *brad@stat.stanford.edu*). This work was supported in part by NIH grant 8R01 EB002784 and by NSF grants DMS 0505673/0804324.

Journal of the American Statistical Association
September 2010, Vol. 105, No. 491, Theory and Methods
DOI: 10.1198/jasa.2010.tm10367

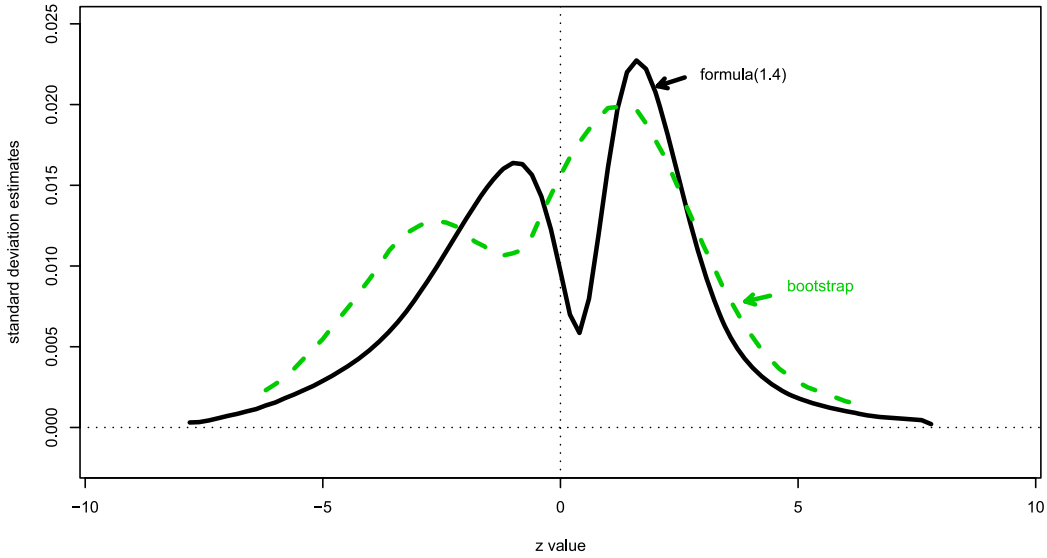

Figure 8. Bootstrap standard deviation estimates for $\hat{F}(x)$, leukemia data, $B = 200$, compared to formula (1.4), Table 1. The online version of this figure is in color.

1 are in the majority, as Schwartzman says, but the upper limit $1+2\alpha$ is well within reach. Here each vector $\mathbf{z}$ has had its mean subtracted. Schwartzman's more general results, not requiring $\alpha_1 = 0$, seem well worth pursuing.

Why have I focused all my attention on normal variates $z_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$? One answer is that z-values are ubiquitous in statistical practice, and originally I wished to develop a single algorithm (`locfdr`) for general use. More to the point, however, is that normality allows one to bring the full power of classical theory—Mehler's identity being *really* classical—to bear, as in Theorem 1. So, as Schwartzman points out, I have reversed my previous habits and begun here with normality, showing only later that the results apply to z-values.

Much of the simultaneous testing literature (though not this paper) considers the identification of "interesting" cases from an ordered set of p-values, $p_{(1)} \le p_{(2)} \le p_{(3)} \le \cdots \le p_{(N)}$. False discovery rate theory uses about the same cutoff point as the Bonferroni bound for the most extreme case,

$$\textit{Fdr}: \quad p_{(1)} \le q/N, \qquad \textit{Bonferroni}: \quad p_{(1)} \le \alpha/N$$

in the usual notation. For the ith ordered case, however, the Fdr cutoff increases to $p_{(i)} \le qi/N$, i times larger than Bonferroni, and in general is usually much more generous than FWER methods in awarding "interesting" status. This generosity has a lot to do with Fdr's popularity. As Westfall points out in his conclusion, Fdr offers fewer advantages in small-N situations, where the interesting values of i may never get very big, making the virtues of permutation-based FWER methods more compelling.

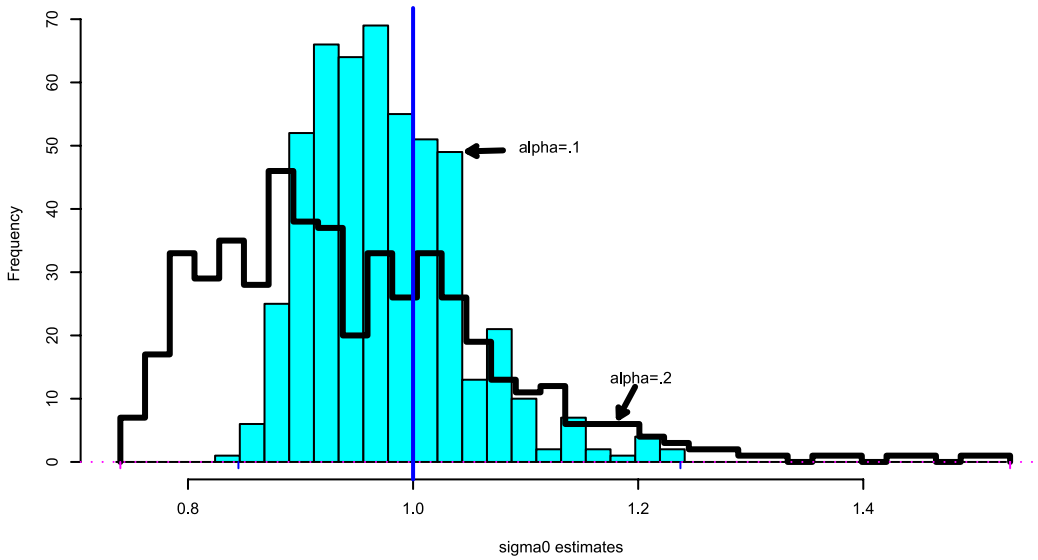

Figure 9. Standard error estimates $\hat{\sigma}_0$ from simulated z-values $z_i \sim \mathcal{N}(0, 1)$ for $i = 1, 2, \ldots, 1000$; rms correlation α equals 0.1 (solid histogram) and 0.2 (line histogram), 500 simulations each. The online version of this figure is in color.

I agree with Cai that validity has been over-emphasized in the Fdr literature, compared to questions of efficiency (i.e., size compared to power). The empirical Bayes interpretation of $\widehat{\text{fdr}}$ or $\widehat{\text{Fdr}}$ as estimates of actual Bayes posterior probabilities is even-handed *via-á-vis* errors of the first and second kind: $\widehat{\text{fdr}}(z)$ estimates the false positive rate and $1 - \widehat{\text{fdr}}(z)$ the true positive rate at every value of z, corresponding to frequentist notions of size and power.

Finally, I am grateful to our editor, Len Stefanski, for featuring this paper and for organizing its discussion at JSM 2010 and in *JASA*.

ADDITIONAL REFERENCE

Benjamini, Y., and Heller, R. (2007), "False Discovery Rates for Spatial Signals," *Journal of the American Statistical Association*, 102, 1272–1281. [1067]