Editorials

False Discovery Rates and the James-Stein Estimator

The new century has brought us a new class of statistics problems, much bigger than their classical counterparts, and often involving thousands of parameters and millions of data points. Happily, it has also brought some powerful new statistical methodologies. The most prominent of these is Benjamini and Hochberg's False Discovery Rate (FDR) procedure, extensively explored in this issue of *Statistica Sinica*, along with ROC techniques. Here, in the brief format of an editorial, I wanted to step back from particulars to take a broader look at the background, virtues, and limitations of FDR methods.

There is an important connection between FDRs and James-Stein estimation. The connection is both historical and methodological, and, perhaps, has something to say about future FDR developments. Suppose we observe $N \geq 3$ independent normal variates z_i, all with variance 1 but having possibly different expectations μ_i,

$$z_i \overset{\text{ind}}{\sim} N(\mu_i, 1), \;\; i = 1, 2, \ldots, N. \tag{1}$$

The obvious estimator of $\mu = (\mu_1, \mu_2, \ldots, \mu_N)'$ is $\hat{\mu}^0 = z = (z_1, z_2, \ldots z_N)'$. James and Stein (1961) proposed the seemingly illogical competitor

$$\hat{\mu}^1 = \left(1 - \frac{N-2}{\sum z_i^2}\right) z, \tag{2}$$

and demonstrated, to the astonishment of the statistics community, that it always dominated $\hat{\mu}^0$ in terms of total expected squared error. Moreover, the improvements could be dramatic in realistic situations.

James and Stein worked in a purely frequentistic framework, but Bayesian connections were soon discovered. Suppose that the parameters μ_i in (1) are themselves normally distributed,

$$\mu_i \overset{\text{ind}}{\sim} N(0, A),\ i = 1, 2, ..., N, \tag{3}$$

where $A > 0$ is some fixed but unknown hyperparameter. Then the Bayes estimator of μ is

$$\hat{\mu}^{\text{Bayes}} = (1 - \frac{1}{A+1})\ z. \tag{4}$$

We don't know the value of 1/(A+1), but an unbiased estimate of it is $(N-2)/(\sum z_i^2)$. Plugging this into (4) gives (2). In other words, the James-Stein rule is an *empirical Bayes* estimator of μ (see Efron and Morris, (1975)).

Benjamini and Hochberg's original 1995 paper concerns testing N null hypotheses on the basis of independently observed p-values "p_i." These can be converted into z-values via $z_i = \Phi^{-1}(p_i)$, with Φ the standard normal cdf. This gets us back to model (1), where the ith null hypothesis is

$$H_{0i} : z_i \sim N(\mu_i, 1) \text{ with } \mu_i = 0. \tag{5}$$

(The alternative $z_i \sim N(\mu_i, 1)$, $\mu_i \neq 0$ is more specialized than necessary, only for the sake of convenient discussion here.)

Let $\widehat{F}(z)$ be the right-sided cdf of the N z-values,

$$\widehat{F}(z) = \#\{z_i \geq z\} / N; \tag{6}$$

$F_0(z) = 1 - \Phi(z)$, the right-sided null hypothesis cdf. Let p_0 be the proportion of true null hypotheses, and then define the estimated FDR to be

$$\widehat{\text{Fdr}}(z) = p_0 F_0(z) / \widehat{F}(z). \tag{7}$$

(The 1995 paper sets p_0 equal to its upper bound 1.) Benjamini and Hochberg's FDR control rule chooses a control level q, say q =.1, computes

$$z_0 = \min_z \left\{ \widehat{\text{Fdr}}(z) \leq q \right\}, \tag{8}$$

and rejects all null hypotheses H_{0i} having $z_i \geq z_0$. Their theorem, almost as surprising as James and Stein's, is that the expected proportion of "false discoveries" produced by this rule; that is, the rejected null hypotheses that are actually true, is less than q.

FDRs underwent the same philosophical progression as the James-Stein estimator — their initial frequentist derivation was followed by a Bayesian reinterpretation. A simple Bayesian "two-groups model" assumes that each of the N cases is either null or non-null with prior probability p_0 or $p_1 = 1-p_0$, and with z-values having right-sided cdf either F_0 or F_1 , where $F_1(z)$ is an unspecified alternative distribution, presumably yielding more extreme z-values than $F_0(z)$ (see Efron (2008)). Letting $F(z)$ be the mixture cdf

$$F(z) = p_0 F_0(z) + p_1 F_1(z), \tag{9}$$

Bayes rule computes the posterior probability of a case being null as

$$\begin{aligned}\mathrm{Fdr}(z) &\equiv \mathrm{Prob}\{\text{case } i \text{ null} \,|\, z_i \geq z\} \\ &= p_0 F_0(z) / F(z)\,.\end{aligned} \tag{10}$$

Comparing (10) with (7) shows that $\widehat{\mathrm{Fdr}}(z)$ is an obvious empirical Bayes estimate of Fdr(z). The control rule (8) amounts to rejecting those cases having sufficiently small estimated posterior probability of being null.

Let $N_z = \#\{z_i \geq z\}$, the number of z_i's exceeding value z. Then $\widehat{F}(z) = N_z / N$ while $F(z) = E\{N_z\} / N.$ All we need for $\widehat{\mathrm{Fdr}}(z)$ to be a useful estimate of Fdr(z) is for N_z to accurately estimate its own expectation. Even if independence of the z_i's fails, as it usually does in microarray applications, $\widehat{\mathrm{Fdr}}(z)$ will still remain nearly unbiased for Fdr(z) (though with increased variability) and we can expect the Benjamini-Hochberg control algorithm to continue functioning reasonably well.

Traditional hypothesis testing techniques aim to control the *probability* of error. FDR methods aim to control an *expectation* that is not a probability (at least not a frequentist one). This represents a major change in the way statisticians do business with the scientific world. FDRs take us away from the familiar territory of p-values and significance levels. I've been pleasantly surprised by how little protest this has aroused from our traditional customers.

The relationship between FDR and the James-Stein estimator can be illustrated in terms of the structural model in (1). Let $g(\mu)$ represent a prior distribution for the expectation parameters μ in the model $z \sim N(\mu,1)$. According to (3),

$$g(\mu) = N(0, A) \tag{11}$$

for James-Stein estimation. FDR methods, as in (5) and (9), take

$$g(\mu) = p_0\delta_0(\mu) + p_1 g_1(\mu), \tag{12}$$

where $\delta_0(\mu)$ is a delta function at 0 representing the null cases $\mu_i = 0$, while $g_1(\mu)$ represents the distribution of non-null cases, those having $\mu_i \neq 0$.

The fact that $g(\mu)$ is much smoother in (11) than (12) hints at estimation difficulties in the FDR context: $\hat{\mu}^1$ is a quite effective approximation to Bayes rule (4) for N as small as 10, while $\widehat{\text{Fdr}}(z)$ requires N on the order of 1,000 to accurately estimate Fdr(z) (Efron and Morris (1975); Efron (2008)). The Benjamini-Hochberg control algorithm "works" for smaller N, in its original frequentist sense, but the diminished empirical Bayes accuracy makes one (me!) worry about the relevance of the FDR control criterion.

There are important intermediate situations lying between (11) and (12). We might, for instance, replace $\delta_0(\mu)$ in (12) with a $N(0,\sigma_0^2)$ distribution, where σ_0 has some small value that allows uninteresting cases to vary a bit from $\mu_i = 0$. That is, we might broaden the definition of "null." In an observational study, σ_0 could represent the disturbing effects of uncontrolled covariates. In this situation the appropriate null distribution, $F_0(z)$ in (7), would be broadened from $N(0,1)$ to $N(0,1+\sigma_0^2)$.

Broadening the null distribution is more than a hypothetical possibility. Figure 1 concerns a microarray experiment involving two types of leukemia. For each of N = 7,128 genes, a two-sample t-statistic "t_i" was computed, comparing type 2 with type 1 patients, and then converted to a z-value "z_i." Under the usual assumptions, we expect $z_i \sim N(0,1)$, the *theoretical null*, to hold for the presumably large majority of genes not involved in type 1/type 2 differences. However the histogram of the 7,128 z-value contradicts this expectation: its center, where the null cases must predominate, is much wider than $N(0,1)$; a normal curve fit to the central 50% of the histogram, the so-called *empirical null*, is $N(0.09,1.68^2)$ (fit by algorithm **locfdr**, Efron (2008)).

Uncontrolled covariation is one of four reasons discussed in Efron (2008) for possible failure of the theoretical null. Whatever the reason, the choice between $N(0,1)$ and $N(0.09,1.68^2)$ makes an enormous difference to an FDR analysis. Figure 2 compares $\widehat{\text{Fdr}}(z)$ curves, slightly smoothed, for the two nulls, showing how much

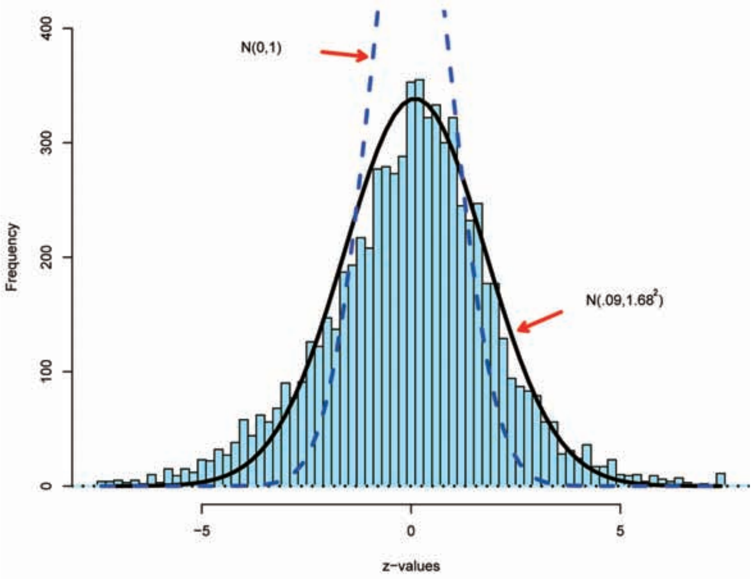

Figure 1: Histogram of z-values comparing two leukemia types, N=7,128 genes; histogram center is highly overdispersed compared to theoretical N(0,1) null distribution; empirical null fit to center is $N(0.09,1.68^2)$. Data are from Golub et al. (1999).

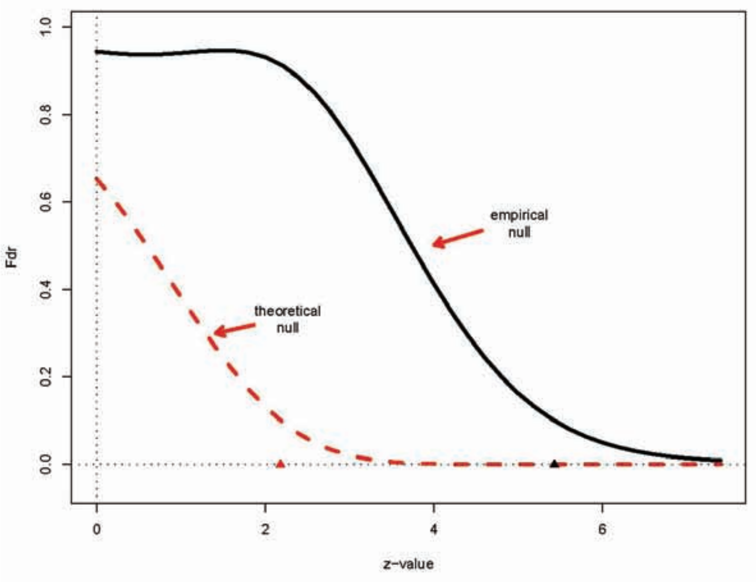

Figure 2: $\widehat{\mathrm{Fdr}}(z)$ for $z > 0$; theoretical null yields 788 genes having $\widehat{\mathrm{Fdr}}(z_i) \le 0.1$, those with $z_i \ge 2.18$; empirical null yields 48, those with $z_i \ge 5.43$.

more conservative is the choice of the empirical null, at least in this case. There is nothing wrong with the FDR algorithm here, except for the tendency in the literature to forget that the numerator $F_0(z)$ in (7) also needs careful consideration.

The charm of exactness exerts a powerful force on statisticians. Both the James-Stein and Benjamini-Hochberg theorems bask in this charm, offering *exact* bounds on their error rates. This is not an unmitigated blessing. The tendency in both literatures has been a further pursuit of exact results. However, this pursuit can come at the expense of ignoring the messy but necessary details that arise in actual applications.

I don't think it is an accident that the James-Stein and Benjamini-Hochberg methods enjoy both frequentist and Bayesian support. Multiple inference, especially large-scale multiple inference, seems to flow easily across the frequentist/Bayes barrier. In addition, the FDR story also combines hypothesis testing with estimation. Perhaps the old categories are breaking down, and FDRs are just the opening salvo in a multi-pronged attack on twenty-first century data analysis problems.

References

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

Efron, B. (2008). Microarrays, empirical Bayes, and the two-groups model. *Statist. Sci.* **23**, 1-47.

Efron, B. and Morris, C. (1975). Data analysis using Stein's estimator and its generalizations. *J. Amer. Statist. Assoc.* **70**, 311-319.

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.

James, W. and Stein, C. (1961). Estimation with quadratic loss. *Proc. Fourth Berkeley Symp. Math. Statist. Probab.* (J. Neyman, ed.) **1**, 361-379. Berkeley, CA: University of California Press

— Bradley Efron