Robbins, Empirical Bayes, and Microarrays

Bradley Efron *

November 1, 2001

Abstract

Empirical Bayes was Herbert Robbins' most influential contribution to statistical theory. It is also an idea of great practical potential. That potential is realized in the analysis of microarrays, a new biogenetic technology for the simultaneous measurement of thousands of gene expression levels.

*Department of Statistics and Division of Biostatistics, Stanford University, Stanford CA 94305; brad@stat.stanford.edu.

1 Introduction

Herbert Robbins ranks high on anyone's list of influential postwar statisticians. Among his many fruitful ideas, Empirical Bayes, which he named as well as developed, has had the biggest effect on statistical thinking.

For reasons that I hope to make clear, current scientific trends favor a greatly increased role for empirical Bayes methods in practical data analysis. This short appreciation is not a review of Robbins' theory. His own review paper in the 1964 Annals of Mathematical Statistics needs no new competition. Rather, after a little bit of history, I will discuss an analysis of microarray data that makes direct use of Robbins' empirical Bayes approach. The suggestion here, made explicit in the final section, is that after 50 years of underuse, we are poised for an avalanche of empirical Bayes applications.

2 Parametric and Nonparametric Empirical Bayes

Table 1 refers to the "missing species problem", a subtle conundrum that has played an important role in empirical Bayes history. The table, abridged from Efron and Thisted (1976), shows the number of distinct words (i.e. words with different spellings, so "tree" and "trees" count separately) appearing in all of Shakespeare's known works, the "canon",

$$n_x = \text{number of distinct words appearing exactly } x \text{ times each.} \tag{2.1}$$

For example $n_1 = 14376$ distinct worlds appeared just once each, $n_2 = 4343$ twice each, etc. The table stops at $x = 30$, but in addition there are 2387 distinct words appearing more than 30 times each, giving a total *observed* Shakespearian vocabulary of

$$\sum_{x \geq 1} n_x = 31534$$

distinct words.

Table 1: Shakespeare's word count frequencies; tabled value n_x is the number of distinct words appearing exactly x times in the Shakespearian canon; 14376 distinct words appear just once, 4343 appear twice each, etc. In addition to those in the table, 2387 distinct words appear more than 30 times each, giving Shakespeare a total of 31534 distinct words appearing in all of his known works.

x:	1	2	3	4	5	6	7	8	9	10
0+	14376	4343	2292	1463	1043	837	638	519	430	364
10+	305	259	242	223	187	181	179	130	127	128
20+	104	105	99	112	93	74	83	76	72	63

The missing species here are the distinct words Shakespeare knew but did not use. By definition there are n_o of them, but of course n_o is missing from Table 1. It doesn't seem possible to learn anything about missing species from Table 1's data, but that is where empirical Bayes thinking comes to the rescue.

Let J denote the true size of Shakespeare's vocabulary (so $J = 31534 + n_o$) and x_j equal the number of times word j appears in the Shakespearian canon; also suppose that the x_j are Poisson random variables each having its own Poisson parameter λ_j,

$$\text{Prob}\{x_j = x\} = e^{-\lambda_j}\lambda_j^x/x! \quad \text{for} \quad x = 0, 1, 2, \ldots . \tag{2.2}$$

Independence of the x_j is not required. Denote the cumulative distribution function (cdf) of the Poisson parameters by $G(\lambda)$,

$$G(\lambda) \equiv \#\{\lambda_j \le \lambda\}/J. \tag{2.3}$$

Now imagine that we could select a word at random with equal probability from the J possibilities. Bayes theorem provides a formula for the expectation of the selected word's λ value given that it occurred x times in the canon, possibly $x = 0$,

$$\begin{aligned} E\{\lambda|x\} &= \frac{\int_0^\infty [e^{-\lambda}\lambda^{x+1}/x!]dG(\lambda)}{\int_0^\infty [e^{-\lambda}\lambda^x/x!]dG(\lambda)} = (x+1)\frac{\int_0^\infty [e^{-\lambda}\lambda^{x+1}/(x+1)!]dG(\lambda)}{\int_0^\infty [e^{-\lambda}\lambda^x/x!]dG(\lambda)} \\ &= (x+1)\frac{\nu_{x+1}}{\nu_x}, \end{aligned} \tag{2.4}$$

where $\nu_x = E_G\{n_x\}$ is the expectation of n_x, (2.1),

$$\nu_x = J\int_0^\infty [e^{-\lambda}\lambda^x/x!]dG(\lambda). \tag{2.5}$$

Substituting n_x for the unobservable ν_x in (2.3) yields what may be the first, and most famous, nonparametric empirical Bayes result,

$$\widehat{E}\{\lambda|x\} = (x+1)\frac{n_{x+1}}{n_x}. \tag{2.6}$$

This formula appears in both Good and Toulmin (1956) (Good credits Turing with some of the ideas) and Robbins (1956). For $x = 1$, Table 1 gives

$$\widehat{E}\{\lambda|x = 2\} = \frac{2 \cdot 4343}{14376} = .604, \tag{2.7}$$

implying that the words appearing once each in the canon are typically over-represented: if we happened to find a collection of previously undiscovered Shakespeare equal in volume to the present canon, the 14376 singleton words would appear an expected total of only $.604 \cdot 14376 = 8623$ times.

Both Robbins and Good & Toulmin applied (2.5) to a variant of the missing species problem. Let

$$r_o = \sum_{x_o=0} \lambda_j/\lambda_+ \quad \left[\lambda_+ = \sum_{j=1}^J \lambda_j\right], \tag{2.8}$$

the proportion of the total expectation in the "missing" class. A good estimate of λ_+ is $N = 884,647$, the total number of words, counting repetitions, in the canon. The numerator of (2.7) is estimated by

$$\widehat{E}\{\lambda|x = 0\} \cdot n_o = \frac{n_1}{n_o} \cdot n_o = n_1, \tag{2.9}$$

so (2.5) yields

$$\widehat{r}_o = n_1/N, \tag{2.10}$$

in our case equaling $14376/884,647 = .016$.

This can be provocatively interpreted as saying that the next "new" word of Shakespeare you find has probability .016 of *not* existing in the current canon. In fact a previously unknown poem was discovered in the Bodelian library in 1985 and attributed by some experts to Shakespeare. Efron & Thisted (1987) applied empirical Bayes results like (2.5) to an analysis of the poem's Shakespearian provenance.

Robbins preferred the term "compound Bayes" for the Shakespeare problem, "empirical Bayes" being reserved for situations where $G(\lambda)$ is a genuine prior distribution rather than just an empirical distribution as in (2.3). This difference was important for the careful decision-theoretic asymptotics in Robbins' pioneering papers, Robbins (1951), Robbins & Hannan (1955). The current climate of more results but less precision tends to ignore the compound-empirical distinction, as I am doing here.

All of this concerns *nonparametric* empirical Bayes. Fisher, Corbet, and Williams (1943) addressed the missing species problem from a parametric empirical Bayes viewpoint. Starting with the Poisson model (2.2), they assumed that $G(\lambda)$ in (2.3) was the cdf of a gamma distribution. The two gamma parameters, scale and index, were then estimated by maximum likelihood applied to the equivalent of Table 1, an early example of hierarchical modeling.

Efron and Morris (1973) appropriated the name "empirical Bayes" for its quintessentially parametric application to Stein estimation. An apology is called for here. We believed that Robbins-type nonparametric empirical Bayes was fundamentally impractical since it required an unimaginably large number of parallel cases to be effective, while Stein's rule applied to as few as three cases at a time. What was unimaginable in 1973 is commonplace in 2001. Nonparametric empirical Bayes applies in an almost off-the-shelf manner to microarrays, the hot new technology that has revolutionized genetic research.

3 Statistical Analysis of Microarrays

Microarrays are devices for measuring gene "expression levels", that is how active a particular gene is in the workings of a given cell. They differ from previous biogenetic technology in being able to measure expression levels for thousands of candidate genes at once. This is a great advantage for microbiologists, speeding the measurement process by a 3 orders of magnitude, but it leads to massive problems of statistical inference, which is where empirical Bayes enters the picture.

Table 2 shows a small part of the data from a microarray experiment concerning stomach cancer: 2640 genes were measured on each of 48 microarrays; each microarray used cells from a different cancer patient, the first 24 patients having less aggressive disease (Type 1) while the second 24 had more aggressive disease (Type 2) giving in total a 2640×48 data matrix of expression levels. The purpose of the study was to discover which genes were more active or less active in Type 2 compared to Type 1 tumors. Newton et al. (2000) gives some background on microarrays, as do Efron, Tibshirani, Storey, and Tusher (2001) and also Efron, Storey, and Tibshirani (2001). Most of the theory that follows is taken directly from the last two references and related work in Tusher et al. (2001), Storey (2001) and Genovese and Wasserman (2001).

If we had only the data for one gene, say gene 1, we could use a two-sample t-statistic to to

Table 2: Some of the expression data for the first 10 genes in the stomach cancer microarray example. There are a total of 2640 genes, each of which had its expression levels measured o 48 microarrays, 24 for Type 1 cancers (less serious) and 24 for Type 2 (more serious); tval is the two-sample t-statistic comparing Type 2 with Type 1; pval is the twosided significance level of tval, 46 degrees of freedom.

type:	1	1 ...	1	1	2	2 ...	2	2	**tval**	**pval**
gene										
1	-0.22	-0.13	-1.23	0.13	-0.80	-0.36	-0.31	0.38	1.550	0.128
2	0.30	-0.12	-0.92	0.02	-1.13	-1.99	0.20	-0.46	2.847	0.007
3	-0.83	-0.01	-0.50	-1.69	-1.89	0.33	-1.12	-0.27	0.850	0.400
4	-0.14	0.69	-0.86	0.27	0.67	1.10	0.42	-0.96	-0.310	0.758
5	0.03	0.25	0.34	0.97	-0.43	0.10	0.03	-1.03	-1.852	0.070
6	0.66	0.68	0.22	0.58	-0.04	-0.09	-0.04	1.11	-2.226	0.031
7	-0.64	-0.36	0.66	0.01	0.18	0.31	0.57	-0.53	0.356	0.723
8	-0.02	-0.15	0.84	-0.13	-0.56	-0.24	-0.39	-0.43	-0.020	0.984
9	0.71	-0.29	0.48	-0.03	-0.56	-0.78	-0.34	0.27	0.460	0.648
10	0.16	-0.04	-0.55	-1.83	-0.90	-0.41	0.56	-0.04	1.914	0.062

test for a difference between the 24 Type 2 measurements versus the 24 Type 1's. Table 2 shows the t value to be 1.550 for gene 1, two-sided p-value 0.128 according to a standard t distribution with 46 degrees of freedom. This isn't significant by the usual .05 standard, but the next gene has $t = 2.847$, $p = .007$, indicating greater expression in Type 2 tumors. Gene 6 is significant in the other direction, indicating greater expression in Type 1. All told, 818 of the 2640 genes were "significant", that is had $p < .05$, but of course we would expect $132 = .05 \cdot 2640$ such cases even if there were no real differences at all. How should the statistician report the results?

Let Y_i be the two-sample t-statistic for gene i. The solid histogram in Figure 1 displays all 2640 Y_i values. We see that the histogram is much wider than the density function $f_o(y)$ for a student's t variate with 46 degrees of freedom. Certainly some of the genes behave differently in Type 2 vis-a-vis Type 1 tumors. Robbins-type empirical Bayes theory will help us quantify the gene by gene evidence for "different behavior".

A very simple Bayesian model assumes that there are two classes of genes. "Different" and "Not Different", meaning that the gene is either differently or not differently expressed in Type 1 and Type 2 tumors. Let the prior probabilities for the two classes be p_1 and $p_o = 1 - p_1$, with corresponding prior densities $f_1(y)$ and $f_o(y)$ for the two-sample t-statistic Y:

$$\begin{aligned} p_1 &= \text{Prob}\{\text{Different}\} & f_1(y) &= \text{density of } Y \text{ if gene "Different"} \\ p_o &= \text{Prob}\{\text{Not Different}\} & f_o(y) &= \text{density of } Y \text{ if gene "Not Different".} \end{aligned} \tag{3.1}$$

Finally, let $f(y)$ be the mixture density

$$f(y) = p_o f_o(y) + p_1 f_1(y). \tag{3.2}$$

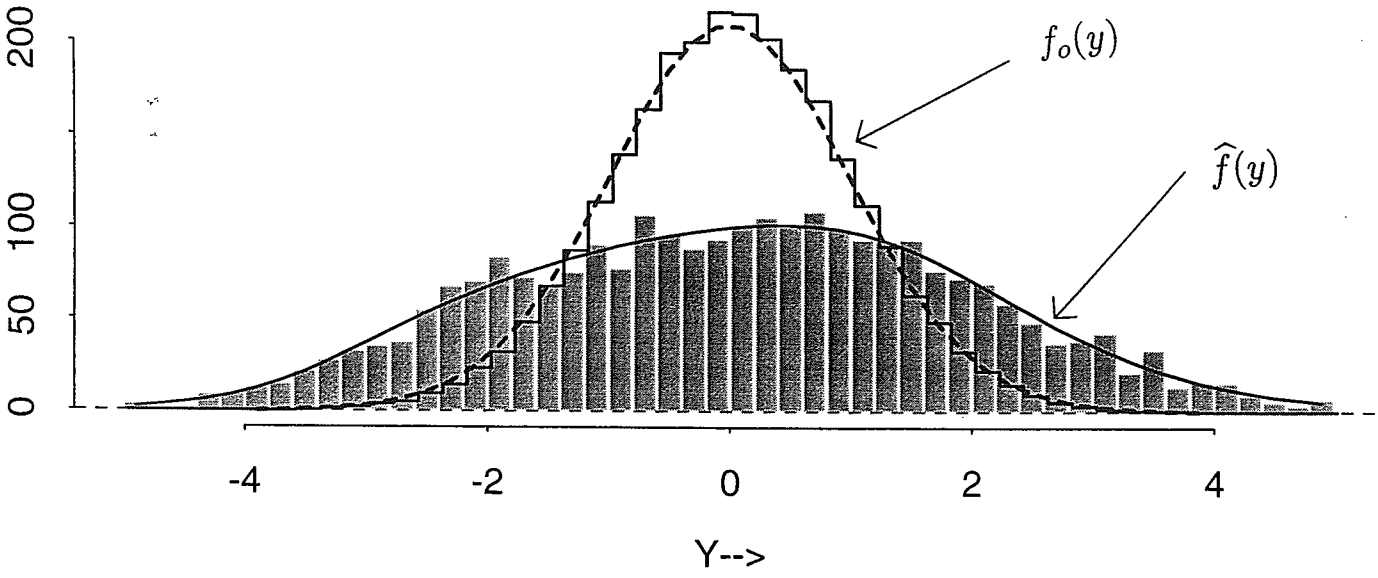

Figure 1: *Solid histogram* shows distribution of the 2640 two-sample t statistics Y_i; this is much wider than density function $f_o(y)$ for a t variate with 46 degrees of freedom, *dashed curve*; *solid curve* $\widehat{f}(y)$ is a smooth fit to the solid histogram; *empty histogram* is permutation estimate of null density f_o, as explained in text. In this case it closely approximates the theoretical null density $f_o(y)$.

We can apply Bayes' theorem to get *a posteriori* probabilities

$$p_1(y) = \text{Prob}\{\text{Different } | Y = y\} = 1 - p_o f_o(y)/f(y) \tag{3.3}$$
$$\text{and}$$
$$p_o(y) = \text{Prob}\{\text{Not Different } | Y = y\} = p_o f_o(y)/f(y).$$

In our case $f_o(y)$ is the standard t density with 46 degrees of freedom. We don't know $f(y)$ but we can estimate it by fitting a smooth curve $\widehat{f}(y)$ to the Y histogram, as in Figure 1, where $\widehat{f}(y)$ is a Poisson GLM fit. This is the crucial empirical Bayes step, analogous to substituting n_x for ν_x in the famous result (2.6). Together these give an estimate of $p_1(y)$, the posterior probability of "Different",

$$\widehat{p}_1(y) = 1 - p_o f_o(y)/\widehat{f}(y) \tag{3.4}$$

The prior probabilities p_1 and $p_o = 1 - p_1$ in (3.1) are unidentifiable without parametric assumptions on the densities $f_1(y)$ and $f_1(y)$. (Robbins assumes normality in the similar example of his 1951 paper.) Figure 2 shows $\widehat{p}_1(y)$, (3.4), for two choices of p_1 : $p_1 = 0$, the most conservative possible choice in terms of minimizing $\widehat{p}_1(y)$; and $p_1 = .531$, the minimum value of p_1 that makes $\widehat{p}_1(y)$ in (3.4) everywhere positive,

$$p_{1,\min} = 1 - \min_y\{\widehat{f}(y)/f_o(y)\} \tag{3.5}$$

This study actually began with more than 10,000 candidate genes, 80% of which were discarded by a rough screening, which helps account for the big value of $p_{1,\min}$.

It might seem that the prior assumption $p_1 = 0$ rules out any interesting posterior probabilities $\widehat{p}_1(Y_i)$, but that is not the case: 389 of the 2640 genes still have $\widehat{p}_1(Y_i) \geq 0.90$. The non-identifiability of p_1 or p_o also shows up in the frequentist multiple-comparison theory discussed in Section 4. It is

the price we pay for using methods that are nonparametric and also *non-structural*: model (3.1)-(3.2) does not require a structural specification for the Y observations (unlike (2.3)-(2.2), where $\lambda \sim G$ and then $x \sim p_o(\lambda)$.)

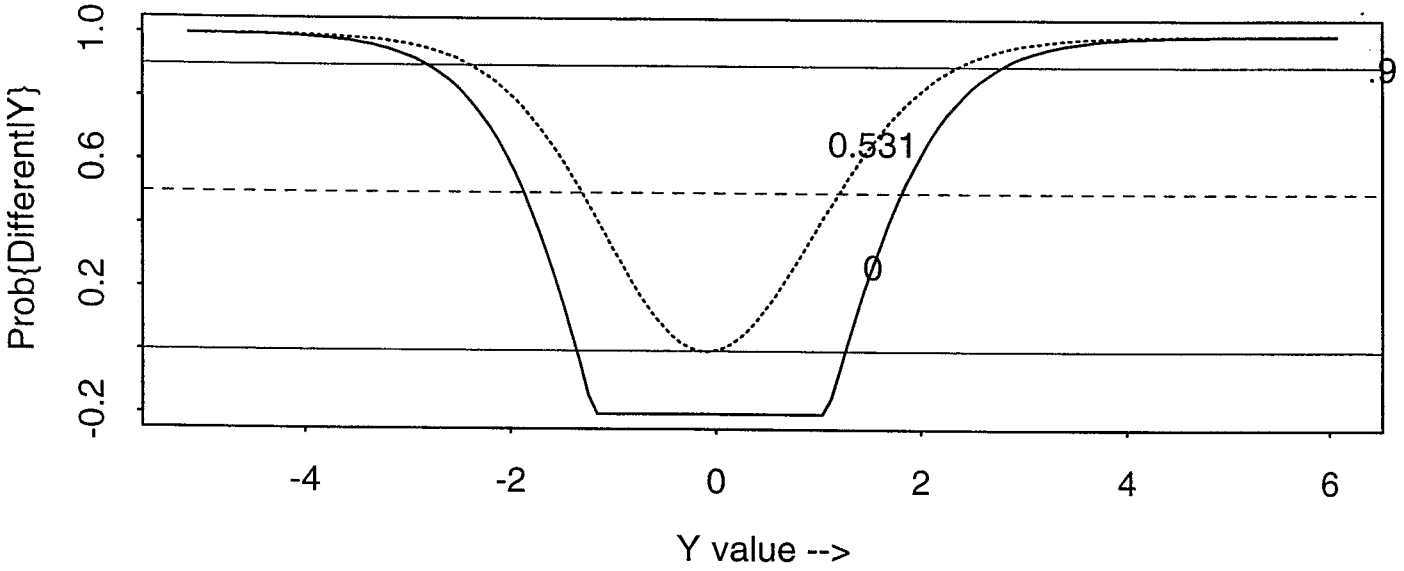

Figure 2: Empirical Bayes posterior probability that a gene is in the "Different" class given t statistic $Y_i = y$, (3.4). *Solid curve* assuming prior probability p_1 of Different equals 0; *Dashed curve* assuming $p_1 = .531$, the smallest value that makes $\widehat{p}_1(y)$ everywhere positive. 389 of the 2640 genes have $\widehat{p}_1(y) \geq .90$, even beginning with the conservative choice $p_1 = 0$.

How do we know that a student's t distribution with 46 degrees of freedom is the appropriate choice for $f_o(y)$ in (3.1)? As a check, permutation methods were used to generate a data-based estimate of f_o: the 48 microarrays were randomly permuted in a balanced way, 12 of the Type 2's moved into the Type 1 class and vice-versa, and the 2640 t-statistics recomputed. This process was independently repeated 20 times. All $20 \cdot 2640$ permutation t values gave the empty histogram in Figure 1, closely following the theoretical t density in this case. Balancing is important here. Unbalanced permutations add a spurious component of variance to the permutated t value, coming from these genes where there is actually a substantial difference between Type 1 and Type 2 response.

Table 3 shows the empirical Bayes estimate $\widehat{p}_1(Y_i)$, (3.4) for the 10 genes of Table 2. Two sets of estimates are given, corresponding to the two curves in Figure 2. Only gene 2 has $\widehat{p}_1(Y_i)$ exceeding .90, but based on prior scientific knowledge the biogeneticist might be interested in genes 6, 10, 5, or even 1. It is important to remember the "empirical" in empirical Bayes. Bootstrap analyses, resampling the microarrays, show some variability in the curves of Figure 2, and considerable variability in the Y_i value for any one gene.

The analysis so far depended on a drastic data reduction, from the full 48-vector $\mathbf{v}_i$ for gene i to the t-statistic Y_i. Information is bound to be lost in the mapping from $\mathbf{v}_i$ to Y_i, but the less we lose the more powerful our analysis, and the better our chance of detecting genuinely Different genes.

One can imagine applying model (3.1), (3.2) directly to the vectors $\mathbf{v}_i$. The theory stays the same, with (3.4) becoming

$$\text{Prob}\{\text{Different}|\mathbf{v}_i\} = 1 - p_o\widehat{f}_o(\mathbf{v}_i)/\widehat{f}(\mathbf{v}_i). \tag{3.6}$$

Table 3: Empirical Bayes estimates of $p_1(y) = \text{Prob}\{\text{Different}|Y\}$ for the 10 genes in Table 2; for $p_1 = 0$ (column 3) and $p_1 = .513$ (column 4). Only gene 2 has $p_1(Y)$ exceeding .90.

gene	tval	pval	$P_o\{\text{Diff}\|Y\}$	$P_{.513}\{\text{Diff}\|Y\}$
1	1.550	0.128	0.289	0.666
2	2.847	0.006	**0.909**	**0.957**
3	0.850	0.400	0.000	0.314
4	-0.310	0.758	0.000	0.023
5	-1.852	0.070	0.487	0.759
6	-2.226	0.030	0.718	0.868
7	0.356	0.724	0.000	0.084
8	-0.020	0.984	0.000	0.002
9	0.460	0.648	0.000	0.124
10	1.914	0.062	0.569	0.798

The trouble comes in trying to estimate the high-dimensional densities $f_o(\mathbf{v})$ and $f(\mathbf{v})$. However we can at least explore various one-dimensional mappings $\mathbf{v}_i \to Y_i$, looking for ones that do not lose much information. Information loss manifests itself by reductions in the likelihood ratio $\widehat{f}_o(Y_i)/\widehat{f}(Y_i)$, which reduces the number of genes having $\text{Prob}\{\text{Different}|Y_i\}$ very large.

Figure 3 compares four choices of the constant a_o in mappings $\mathbf{v}_i \to Y_i$ of the form

$$\mathbf{v}_i \to \text{num}_i/(a_o + \text{den}_i), \tag{3.7}$$

where num_i and den_i are the numerator and denominator of the usual two-sample t-statistic: $a_o = 0$ gives the usual t-statistic; $a_o \to \infty$ makes Y_i proportional to num_i; "$a_o = .5$" and "$a_o = .9$" corresponds to intermediate cases where a_o is taken to be the 50th, or 90th, percentile of all 2640 den_i values. The choice "$a_o = .9$" gave the best results in the experiment featured in Efron, Tibshirani, Storey, and Tusher (2001).

Figure 3 compares $\widehat{\text{P}}\text{rob}\{\text{Different}|Y\}$ for the four mappings, always taking $p_1 = 0$, $p_o = 1$ in (3.4), corresponding to the solid curve in Figure 2. Here the vertical axis has been transformed to the logit scale, to emphasize difference in the tails, while for each mapping the Y_i values have been monotonically transformed to have a Normal$(0, 1)$ empirical distribution. We can see that the choice $a_o = \text{inf}$ is bad in this case, indicating very few "Different" genes. Overall, our original choice $a_o = 0$ performs best.

The important practical point is that microarray data sets are large enough to support a lot of numerical experimentation. We can be quite empirical in our empirical Bayes analysis, avoiding arbitrary *a priori* modeling in favor of data-based investigation. Efron, Tibshirani, Tusher (2001) and Efron, Storey and Tibshirani (2001) discuss the investigative possibilities and pitfalls, including the tacit exchangeability assumptions we have been making, see Section 4.

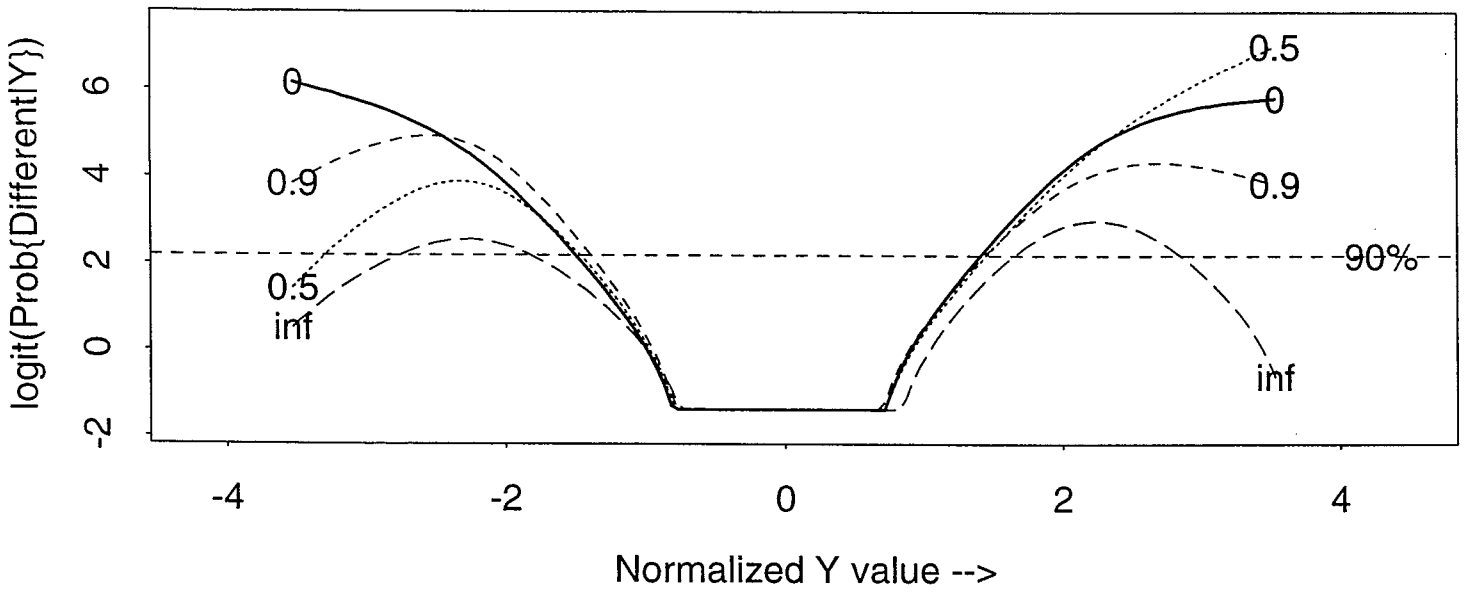

Figure 3: Prob{Different|Y} for four different choices of the constant a_o in the summary statistic Y, (3.7); choice $a_o = 0$, the t-statistic, gives best overall results; $a_o = \infty$, the difference of the type means, is worst. As in Figure 2 except that the vertical axis have been transformed to the logit scale, while the Y_i have been normalized.

4 Empirical Bayes and False Discovery Rates

The Robbins-type empirical Bayes analysis of Section 3 is closely related to Benjamini and Hochberg's (1995) frequentist theory of False Discovery Rates, a promising new multiple comparison criterion. This relationship raises the hope, perhaps illusory, of improving the connection between Bayesian and frequentist testing theory.

Here is a brief description of the FDR theory as it applies to the situation of Figure 1. Let H_i be the null hypothesis that gene i is in the "Not Different" class, in which case Y_i, the two-sample t-statistic has student's t distribution with 46 degrees of freedom,

$$H_i : \; Y_i \sim t_{46}. \tag{4.1}$$

The p value for testing H_i against the alternative that gene i is *less* expressed for Type 2 than for Type 1 tumors is

$$P_i = \text{Prob}\{t_{46} \le Y_i\}. \tag{4.2}$$

(Of course we could just as well test in the other direction.)

Letting $P_{(1)} \le P_{(2)} \cdots \le P_{(n)}$, $n = 2640$, Benjamini and Hochberg consider the following simultaneous testing rule: for a fixed choice of α between 0 and 1 define

$$i_\alpha = \underset{i}{\text{argmax}}\left\{P_{(i)} \le \frac{i}{n}\,\frac{\alpha}{p_o}\right\} \qquad [p_o = \text{proportion of true } H_i], \tag{4.3}$$

and reject all H_i with $P_i \le P_{(i_\alpha)}$. They then show that the False Discovery Rate of this rule,

$$\text{FDR} = E\{\text{proportion of rejected } H_i \text{ that are actually true}\}, \tag{4.4}$$

is bounded above by α. Their 1995 paper assumed independence of the Y_i, but recent work has substantially relaxed this assumption, Benjamini and Yekutieli (2001).

Notice that p_o in (4.3) has nearly the same definition as in the empirical Bayes setup (3.1). As before, p_o is unknown and unidentifiable, but the most conservative choice $p_o = 1$ can be used, just as in Section 3. Figure 4 graphically demonstrates that $i_\alpha = 158$ for the cancer data, with $\alpha = .05$ and $p_o = 1$, so that the FDR rule rejects H_i for the 158 genes having $P_i \leq P_{(158)} = .00298$. Benjamini and Hochberg's theorem says that we expect no more than $7.9 = .05 \cdot 158$ of the 158 rejected H_i to actually be true.

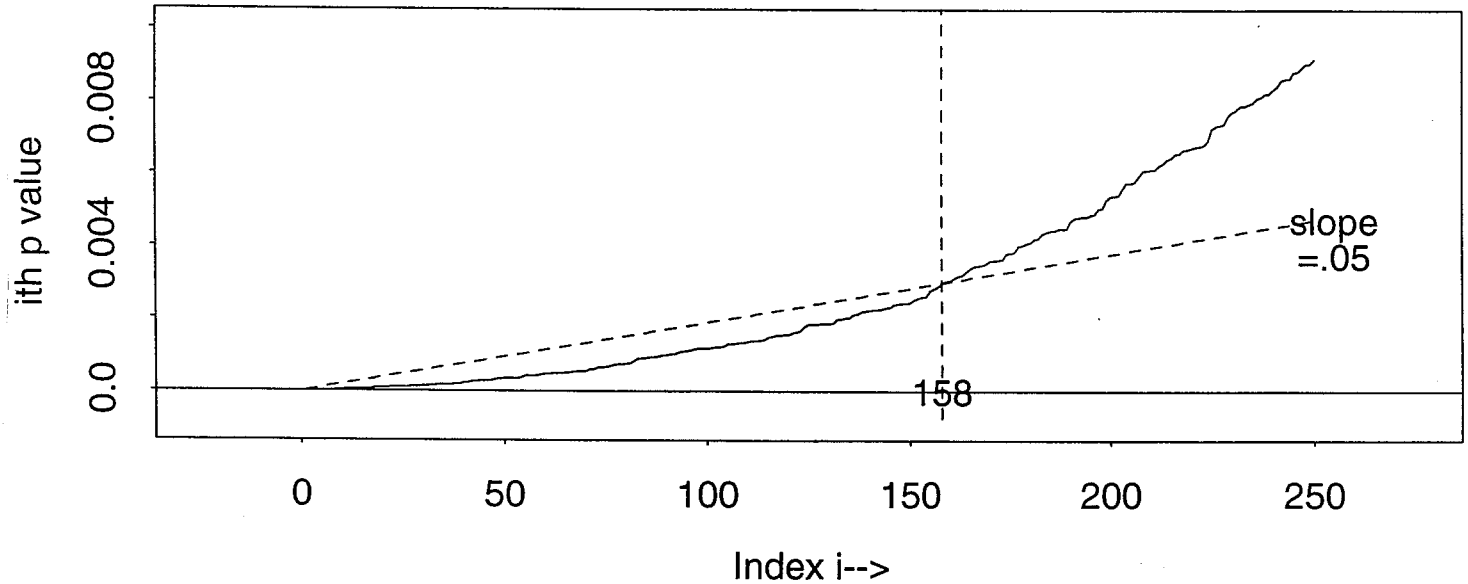

Figure 4: Plot of $P_{(i)}$ vs i, (4.2), for the 2460 genes, $i = 1, 2, \ldots, 250$; the FDR rule (4.3) with $\alpha = .05$, $p_o = 1$, rejects H_i for $P_i \leq P_{(158)} = .00298$.

The close connection between False Discovery Rates and the empirical Bayes methodology of Section 3 follows directly from Bayes theorem. Let $F_o(y)$ and $F(y)$ be the cdfs corresponding to $f_o(y)$ and $f(y)$ in (3.1), and define the "Bayesian FDR" for the rejection rule $\{Y_i \leq y\}$ to be

$$\begin{aligned} \mathrm{Fdr}(y) &\equiv p_o F_o(y)/F(y) \\ &= \mathrm{Prob}\{\text{gene } i \text{ Not Different} | Y_i \leq y\} \end{aligned} \tag{4.5}$$

(called the "q-value" in Storey, 2001.) If there are say N_y genes having $Y_i \leq y$ then among these the expected number of Not Different genes is $N_y \cdot \mathrm{Fdr}(y)$. This justifies calling $\mathrm{Fdr}(y)$ the Bayesian false discovery rate.

The obvious nonparametric estimate for $\mathrm{Fdr}(y)$ is

$$\widehat{\mathrm{Fdr}}(y) = p_o F_o(y)/\widehat{F}(y), \tag{4.6}$$

where $\widehat{F}(y)$ is the usual empirical cdf of the Y_i. Then it is easy, as in Efron, Storey and Tibshirani (2001), to prove the following result:

Equivalence Theorem For given α and p_o, the Benjamini-Hochberg rule is equivalent to rejecting all H_i having $Y_i \leq y_\alpha$, where

$$y_\alpha = \max_y \{\widehat{\mathrm{Fdr}}(y) \leq \alpha\} \tag{4.7}$$

The equivalence theorem makes an important connection between empirical Bayes and frequentist testing criteria: if we chose the rejection region $\{Y_i \leq y\}$ as large as possible subject to the constraint that the estimated Bayes proportion of false discoveries is less than α, than the frequentist expected proportion of false discoveries is also less than α. One can simultaneously be a Bayesian and a frequentist in this case, usually a good sign for both methodologies.

Tail area rejection regions like $\{Y_i \leq y\}$ are natural in the frequentist framework. The empirical Bayes approach suggests a local version of FDR,

$$\text{fdr}(y) = p_o f_o(y)/f(y) = \text{Prob}\{\text{gene } i \text{ Not Different}|Y_i = y\}. \tag{4.8}$$

Consider a small interval on the Y axis, for example $\mathcal{Y} = [-3, -2.8]$. We expect about

$$p_o \cdot n \cdot f_o(2.9) \cdot 0.2 = p_o \cdot 4.05$$

"Not Different" genes in $\mathcal{Y}$ under model (3.1). Actually we observed 36 genes in $\mathcal{Y}$, giving

$$\widehat{f}(y) = \frac{36}{n \cdot 0.2}$$

as the nonparametric empirical estimate of $f(y)$. The corresponding local fdr estimate is

$$\widehat{\text{fdr}}(y) = p_o \frac{4.05}{36} = p_o \cdot .113. \tag{4.9}$$

The conservative choice $p_o = 1$ gives $\widehat{\text{fdr}}(y) = .113$, or equivalently at (3.4), $\widehat{p}_1(y) = 1 - .113 = .887$. This makes the empirical Bayes statement (3.4) almost obvious; we observe 36 genes in $\mathcal{Y}$, and expect only about 4 of these to be "Not Different". Therefore we believe that about 8/9 of the 36 are in the "Different" class. Here the exchangeability assumptions underlying Robbins-type analyses are apparent: the 36 genes must be *a priori* interchangeable to justify believing $\widehat{p}_1(y) = .887$ for any one of them. Section 4.3 of Efron, Storey, and Tibshirani (2001) modifies (3.4) to handle the case of varying *a priori* beliefs.

5 Conclusion

The period between 1945 and 1980, when Robbins was most active, witnessed a host of new methodological developments: nonparametrics, robustness, Kaplan-Meier, proportional hazards, bootstrap, jackknife, Markov chain Monte Carlo, all depending at some level on advances in computation. What was *not* happening was the advancement of basic statistical theory. A Karl Pearson or Gosset dropped into the 21st century would be impressed with our technology but familiar with most of the underlying principles.

Empirical Bayes, both of the Stein and Robbins variety, is the great exception. It is definitely post Fisher-Neyman-Wald in spirit, pointing the way toward an unexpected synthesis of Bayesian and frequentist points of view. Moreover, it is a practical advance as well as a theoretical one. It is possible for empirical Bayes methods to reduce the risk of their classical competitors factors of 2 or more, as shown by the examples in Efron and Morris (1975).

Why hasn't there been a landrush of empirical Bayes applications? The obvious answer is that scientists have not brought us many data sets having the parallel structure necessary for empirical Bayes to do its stuff. This begs the question. Statisticians are more than just passive processors

of whatever problems happen to come our way. Fisher's theory of efficient experimental design greatly influenced the form of 20th century data sets. Analysis of variance fits an amazing number of situations, but that's at least partly because research scientists know we can effectively analyze ANOVA data.

If statisticians demonstrate efficient ways of analyzing parallel data then we will start seeing more parallelism in data base design. Microarrays and their connection with Robbins-type empirical Bayes analysis are an emphatic case in point. There seems to be a good chance that Robbins was 50 years ahead of his time, and that a statistical theory of the 1950's will shine in the 21st century.

Acknowledgment

This research was supported in part by National Institute of Health grant 2R01 CA59039 and by National Science Foundation grant DMS-0072360.

References

Benjamini, Y. and Hochberg, Y. (1995), "Controlling the False Discovery Rate: a practical and powerful approach to multiple testing". *Jour. Royal Stat. Soc., B* **57**, 289-300.

Benjamini, Y. and Yekutieli, Y. (2001), "The control of the False Discovery Rate in multiple testing under dependency", to appear *Ann. Stat.*

Efron, B. and Morris, C. (1973), "Stein's estimation rule and its competitors – an empirical Bayes approach". *JASA* **68**, 117-30.

Efron, B. and Morris, C. (1975), "Data analysis using Stein's estimator and its generalizations". *JASA* **70**, 311-19.

Efron, B. and Thisted, R. (1976), "Estimating the number of unseen species: how many words did Shakespeare know?" *Biometrika* **63**, 435-7.

Efron, B. and Thisted, R. (1987), "Did Shakespeare write a newly-discovered poem?" *Biometrika* **74**, 445-55.

Efron, B., Storey, J. and Tibshirani, R. (2001), "Microarrays, empirical Bayes methods, and False Discovery Rates". Stanford Technical Report, brad@stat.stanford.edu.

Efron, B., Tibshirani, R., Storey, J., and Tusher, V. (2001), "Empirical Bayes analysis of a microarray experiment", to appear in *JASA*.

Fisher, R., Corbet, A., and Williams, C. (1983), "The relation between the number of species and the number of individuals in a random sample of an animal population". *Jour. Animal Ecology* **12**, 42-58.

Genovese, C. and Wasserman, L. (2001), "Operating characteristics and extensions of the FDR procedure". Carnegie Mellon Technical Report.

Good, I. and Toulmin, G. (1956), "The number of new species and the increase in population coverage when a sample is increased". *Biometrika* **43**, 45-63.

Robbins, H. (1951), "Asymptotically sub-minimax solutions of compound statistical decision problems". *Proc. Second Berkeley Symp.* **1**, 131-148, Univ. Calif. Press.

Robbins, H. and Hannan, J. (1955), "Asymptotic solutions of the compound decision problem for two completely specified distributions". *Ann. Math. Stat.* **26**, 37-51.

Robbins, H. (1956), "An empirical Bayes approach to statistics". *Proc. Third Berkeley Symp.* **1**, 152-163, Univ. Calif. Press.

Robbins, H. (1964), "The empirical Bayes approach to statistical decision problems". *Annals Math. Stat.* **35**, 1-20.

Storey, J. (2001), "The False Discovery Rate: a Bayesian interpretation and the q-value". Stanford Technical Report, jstorey@stat.stanford.edu.

Tusher, V., Tibshirani, R., and Chu, G. (2001), "Significance analysis of microarrays applied to transcriptional responses to ionizing radiation". *PNAS* **98**, 5116-21.