An Introduction to Empirical Bayes Data Analysis

GEORGE CASELLA*

Empirical Bayes methods have been shown to be powerful data-analysis tools in recent years. The empirical Bayes model is much richer than either the classical or the ordinary Bayes model and often provides superior estimates of parameters. An introduction to some empirical Bayes methods is given, and these methods are illustrated with two examples.

KEY WORDS: Stein estimation; Normal distribution; Binomial distribution.

1. INTRODUCTION

Empirical Bayes methods have been around for quite a long time. Their roots can be traced back to work by von Mises in the 1940s (see Maritz 1970), but the first major work must be attributed to Robbins (1955), although his formulation is somewhat different from that used here. One might refer to Robbins's formulation as nonparametric empirical Bayes, whereas the formulation discussed here can be referred to as parametric empirical Bayes. The major difference is that the parametric approach specifies a parametric family of prior distributions, but the nonparametric approach leaves the prior completely unspecified. We will deal here only with parametric empirical Bayes methods and will refer to them simply as empirical Bayes methods.

Although the idea of a parametric empirical Bayes analysis is not new, the first major work in this area did not appear until the early 1970s in a series of papers by Efron and Morris (1972, 1973, 1975), and one might rightfully say that they are the founders of modern empirical Bayes data analysis. Efron and Morris (1977) is an excellent, fairly nontechnical account of the interrelationship between these methods and the so-called Stein effect.

Empirical Bayes methods have become increasingly popular and have been applied to many types of problems. Some examples are fire alarm probabilities (Carter and Rolph 1974), revenue sharing (Fay and Herriot 1979), quality assurance (Hoadley 1981), and law school admissions (Rubin 1981). More recently, Morris (1983) formulated a theory of parametric empirical Bayes inference.

The purpose here is to give a simple introduction to empirical Bayes methods and illustrate them with two examples.

2. EMPIRICAL BAYES ESTIMATORS FOR THE NORMAL CASE

Suppose that we observe p random variables, each from a normal population with different means but the same known variance, that is,

$$X_i \sim n(\theta_i, \sigma^2), \qquad i = 1, \ldots, p. \tag{2.1}$$

(Think of a balanced one-way analysis of variance, with the X_i representing the cell means.) The cases of unknown variance or different sample sizes per cell can also be handled, but here we will stay with this simple case.

The usual, or classical, estimator of θ_i is X_i, the observation (or cell mean). This estimator has many optimality properties (best linear unbiased, maximum likelihood, minimax, etc.), but we can do better.

For the moment, make the Bayesian assumption that

$$\theta_i \sim n(\mu, \tau^2), \qquad i = 1, \ldots, p. \tag{2.2}$$

The Bayes estimate for θ_i, $\delta^B(X_i)$, is given by

$$\delta^B(X_i) = [\sigma^2/(\sigma^2 + \tau^2)]\mu + [\tau^2/(\tau^2 + \sigma^2)]X_i. \tag{2.3}$$

Note that $\delta^B(X_i)$ is a weighted average of μ (the prior estimate) and X_i (the sample estimate). The weights used in the weighted average depend on the relative sizes of τ^2 (the prior variance) and σ^2 (the sample variance). As τ^2/σ^2 gets smaller, more weight is put on μ. Thus the relative accuracy of the estimates X_i and μ determines how much weight they receive in the weighted average.

$\delta^B(X_i)$ is the Bayes estimate because it is the mean of the posterior distribution, the distribution of θ_i given X_i, denoted by $\pi(\theta_i|X_i)$. A standard calculation shows that

$$\pi(\theta_i|X_i) \sim n[\delta^B(X_i), \sigma^2\tau^2/(\sigma^2 + \tau^2)], \qquad i = 1, \ldots, p. \tag{2.4}$$

The empirical Bayesian agrees with the Bayes model but refuses to specify values for μ and τ^2. Instead, he estimates these parameters from the data. All of the information about μ and τ^2 is contained in the marginal distribution of X_i (unconditional on θ_i), and another standard calculation shows that this marginal distribution, $f(X_i)$, is given by

$$f(X_i) \sim n(\mu, \sigma^2 + \tau^2), \qquad i = 1, \ldots, p. \tag{2.5}$$

Thus unconditionally, we can regard the X_i's as coming from the same population. This assumption was already implicit in the Bayes model, since each θ_i had the same prior distribution. In many cases this assumption is also quite reasonable—think of a one-way analysis of variance in which the treatments are defined by levels of a particular factor. It is reasonable to assume that there is some distant, underlying similarity in the responses.

Using (2.5), we can construct estimates of the Bayes quantities in (2.3). In particular, we have

$$E(\bar{X}) = \mu, \; E\left[\frac{(p-3)\sigma^2}{\Sigma(X_i - \bar{X})^2}\right] = \frac{\sigma^2}{\sigma^2 + \tau^2}, \tag{2.6}$$

where the expectation is taken over the marginal distribution of the X_i's. From (2.6), we have unbiased estimators of the

*George Casella is Associate Professor, Biometrics Unit, Cornell University, Ithaca, NY 14853. This research was supported by National Science Foundation Grant MCS81-02541.

Bayes quantities in (2.3), and we can construct an empirical Bayes estimator of θ_i by replacing these quantities by their estimates. Thus an empirical Bayes estimator of θ_i, $\delta_i^E(X)$, is given by

$$\delta_i^E(X) = \left[\frac{(p-3)\sigma^2}{\Sigma(X_i - \bar{X})^2}\right]\bar{X} + \left[1 - \frac{(p-3)\sigma^2}{\Sigma(X_i - \bar{X})^2}\right]X_i. \quad (2.7)$$

Note that δ_i^E uses information from all of the X_i's when estimating each θ_i. This takes advantage of what has come to be known as the Stein effect (e.g., see Stein 1981 or Berger 1982). Simply put, the Stein effect asserts that estimates can be improved by using information from all coordinates when estimating each coordinate.

The empirical Bayes estimator $\delta_i^E(X)$ is a good estimator of θ_i. We will see later how it performs on data, but it also has an extremely appealing theoretical property: on the average, it is always closer to θ_i than X_i. We can measure the worth of an estimator δ_i by considering $\Sigma(\theta_i - \delta_i)^2$, the sum of the squared differences between the estimator and the parameter. If $p \geq 4$, it is true that

$$E\left\{\sum_{i=1}^{p}[\theta_i - \delta_i^E(X)]^2\right\} < E\left[\sum_{i=1}^{p}(\theta_i - X_i)^2\right], \quad \text{for all } \theta_i, \quad (2.8)$$

where the expectation here is over the distribution of X_i given θ_i, $X_i \sim n(\theta_i, \sigma^2)$. In this sense, $\delta_i^E(X)$ is always closer to θ_i than X_i. [For a rigorous proof of (2.8), see Efron and Morris 1973.]

The quantities in (2.8) are called the mean squared error (MSE) of the respective estimators and are functions of θ_i and σ^2 only through the quantity $\Sigma_{i=1}^p \theta_i^2/\sigma^2$. On examining Figure 1, it is fairly obvious that the empirical Bayes estimator has the most desirable MSE.

3. SOME EMPIRICAL BAYES INTUITION

There is a nice intuitive justification of the empirical Bayes estimator of (2.7) in the one-way analysis of variance (ANOVA). Suppose that there are five treatments. Let $X_1, \ldots, X_5$ represent observed cell means and $\theta_1, \ldots, \theta_5$ represent true cell means. The ANOVA F statistic tests the hypotheses

$$H_0: \text{all } \theta_i\text{'s equal} \quad \text{vs.} \quad H_A: \text{not } H_0. \quad (3.1)$$

We can regard these hypotheses as two extremes: If H_0 is true, then we should estimate each θ_i with $\bar{X} = \Sigma X_i/5$ (since all of the θ_i's are equal), whereas if H_A is true, then we should estimate each θ_i with X_i. The empirical Bayes estimator, given in (2.7), is a compromise between these two extremes, as can be seen in Figure 2.

Note how the empirical Bayes estimator affects the extreme means (X_1 and X_5) much more than it affects the means that are close to $\bar{X}$. In most cases this type of shrinkage will improve the estimate of θ_i: the extreme cell means are often overestimates or underestimates. One might say that the empirical Bayes estimator anticipates regression to the mean.

The amount of shrinkage in the empirical Bayes estimator is directly related to the F statistic that tests the ANOVA null hypothesis. If there are p treatments, the F statistic is

$$F = [\Sigma(X_i - \bar{X})^2/(p-1)]/\hat{\sigma}^2, \quad (3.2)$$

where $\hat{\sigma}^2$ estimates σ^2. Since we are dealing here with known σ^2, the ANOVA null hypothesis would be tested by

$$T = [\Sigma(X_i - \bar{X})^2/(p-1)]/\sigma^2 \sim \chi^2_{p-1}/(p-1), \quad (3.3)$$

and large values of T would lead to rejection of H_0: all θ_i's equal. By using (3.3), the empirical Bayes estimator of (2.7) can be written as

$$\delta_i^E(X_i) = \left(\frac{p-3}{p-1}\right)T^{-1}\bar{X} + \left[1 - \left(\frac{p-3}{p-1}\right)T^{-1}\right]X_i. \quad (3.4)$$

As T becomes large (and the data support H_A), $\delta_i^E(X_i)$ puts more weight on X_i and less on $\bar{X}$. Thus $\delta^E(X_i)$ puts more weight on the estimate (X_i or $\bar{X}$) that seems most reasonable based on the evidence from all of the data.

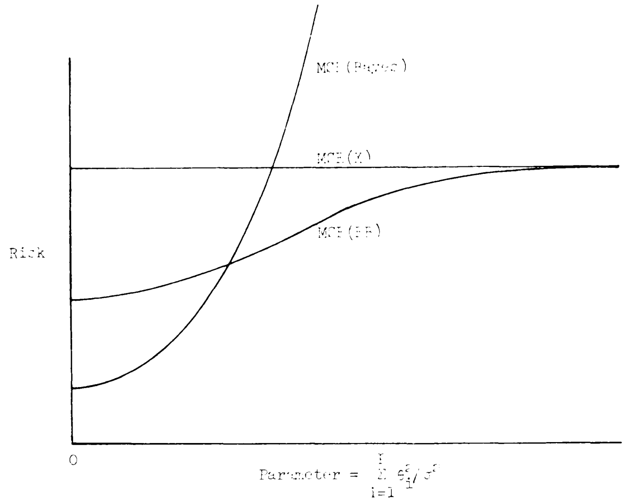

Figure 1. MSE of the Usual Estimator, X, the Bayes Estimator, δ^B, and the Empirical Bayes (EB) Estimator, δ^E.

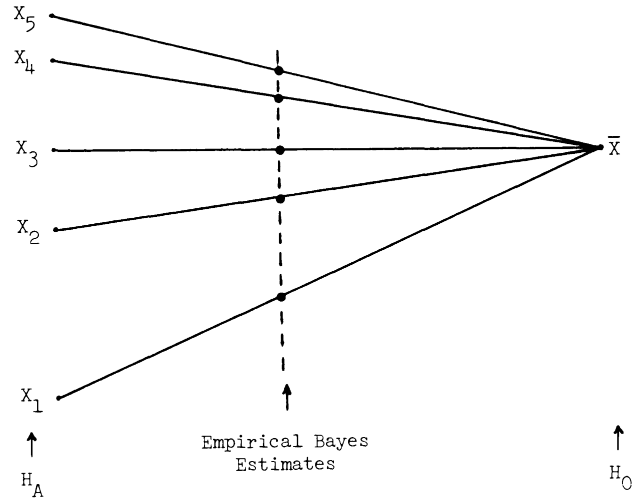

Figure 2. The Empirical Bayes Estimator in the One-Way ANOVA.

4. EXAMPLES OF EMPIRICAL BAYES ESTIMATES

The following two examples were chosen because, in both cases, the parameter values were available. Thus it is possible to assess directly the performance of the estimators.

Example 1. Estimating Batting Averages. Efron and Morris (1975) reported the batting averages of 18 major league baseball players after their first 45 at bats. The problem is to estimate their final batting average. For simplicity, we will only consider here a subset of their data, consisting of seven players selected to be illustrative. (The highest-ranked, the lowest-ranked, and five other players, chosen at random, were included.)

It is reasonable to assume that each time at bat is a binomial trial, with the probability of success equal to the player's true batting average. With 45 trials, the normal approximation seems reasonable. [Actually, the arc sine square root transformation was performed on the data, which were then recentered to resemble batting averages. The variance attached to each player's observed average is $(.0659)^2$.]

Thus we can model each observed batting average, X_i, by

$$X_i \sim n(\theta_i, \sigma^2), \tag{4.1}$$

where θ_i = true batting average and $\sigma^2 = (.0659)^2$. We then use the Bayes prior, $\theta_i \sim n(\mu, \tau^2)$, and construct the empirical Bayes estimator as indicated in Section 2. The data, calculations, and final batting averages (true θ_i) are given in Table 1.

The empirical Bayes estimators are closer to the θ_i's than the classical estimators, the X_i's. The improvement in MSE is remarkable: $.355/1.084 = .327$, meaning a 67% reduction in MSE. [Here we have scaled the MSE, so for example, $1.084 = \Sigma(X_i - \theta_i)^2/7\sigma^2$. Of course, this does not affect the comparisons with the empirical Bayes estimator.]

The empirical Bayes estimator performed well because it anticipated regression toward the mean. The player who was batting .395 after 45 at bats was doing unusually well (playing above his head), and it would be unreasonable to expect him to continue at such a pace. Notice also that both X_i and δ_i^E failed miserably with player 7, who had (for him) an unusually poor start. (An explanation for this failure may be the fact that player 7 was Thurmon Munson, and these data were taken in his rookie year. Munson went on to become a consistently excellent ball player.)

Table 1. Baseball Data

Player	X_i (observed batting average)	θ_i (final batting average)	$\delta_i^E(X)$ (empirical Bayes estimate)
1	.395	.346	.341
2	.355	.279	.321
3	.313	.276	.300
4	.291	.266	.289
5	.247	.271	.266
6	.224	.266	.255
7	.175	.318	.230
MSE	1.084		.355

NOTE: $\bar{X} = .286$; $\Sigma(X_i - \bar{X})^2 = .035$; $4\sigma^2/\Sigma(X_i - \bar{X})^2 = .495$; and $\delta_i^E(X) = (.495)(.286) + .505X_i = .142 + .505X_i$.

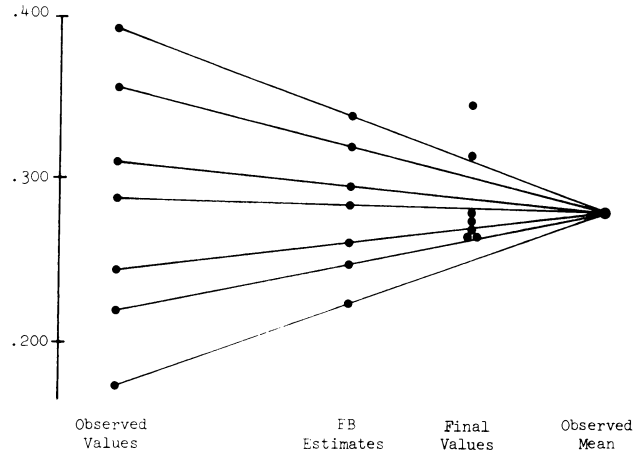

Figure 3. Graphical Display of the Baseball Data.

A graphical display, such as the one in Figure 2, serves to support further the claim that regression toward the mean is a very real effect. Examining Figure 3, and noting how close together the θ_i's are (compared to the X_i's and even the δ_i^E's), reveals that the empirical Bayes estimates are vastly superior to the usual ones.

Example 2. Assessing Consumer Intent. This example was chosen not only because the parameters were available but also to illustrate the empirical Bayes technique for distributions other than the normal distribution. The data were taken from Juster (1966) and were also analyzed by Morrison (1979), using techniques outlined by Sutherland et al. (1975). In fact, Morrison used some highly sophisticated empirical Bayes techniques and obtained even better estimates than those presented here.

The problem here is to estimate the probability that a consumer will purchase a given product, given his stated probability (intent) of such an event. Here we will concentrate on only a portion of Juster's data, in which 447 randomly selected people were asked this question: Taking everything into account, what are the prospects that you or some member of your family will buy a car sometime during the next 12 months? The prospects were ranked on a scale from 0 to 1: Certain (10 in 10), Almost Sure (9 in 10), Very Probably (8 in 10), . . ., Very Slight Possibility (1 in 10), and No Chance (0 in 10). The distribution of responses is given in Table 2.

The data were grouped by Juster (as indicated in Table 2) in order to increase the sample sizes. The weighted averages of these groups are also given in Table 2. This grouping was also used by Morrison (1979) and will be used here. Thus we are dealing with five intent groups.

Before proceeding to a formal model, it ought to be observed that these data should almost certainly be shrunk toward their mean. It is quite unreasonable to assume that

Table 2. Consumer Intent Data

	Intent										
	0	.1	.2	.3	.4	.5	.6	.7	.8	.9	1
Responses	293	26	21	21	10	9	12	13	11	10	21
Weighted averages			.19			.51			.79		

none of the 293 people in the 0-intent group will buy a car; thus 0 is certainly an underestimate of the intent. The same type of argument applies to the group with intent = 1.

The model for these data, used by Morrison and others, is that I_i^S, the stated intent of person i, can be modeled as a binomial random variable with $n = 10$ and $p = I_i^T$, the true intent. That is,

$$I_i^S \sim \text{binomial}(10, I_i^T). \tag{4.2}$$

The justification for this model is that an individual with true intent I^T responds binomially—0 or 1—in an independent fashion to each point on the intent scale with probabilities I^T and $1 - I^T$, respectively. The stated intention is then the sum of these 0, 1 responses.

Although this model may sound strange, it has been widely used and justified in both marketing and psychology literature (Morrison 1979). From a practical point of view, it also seems to work well.

There is a minor problem with scaling, in that the stated intention is on a 0–1 scale and the modeled intention is on a 0–10 scale. This can, of course, be handled rather easily, and we will not go into such details here.

The empirical Bayes model also specifies that

$$I_i^T \sim \text{beta}(\alpha, \beta); \tag{4.3}$$

that is, the true intentions are drawn from a beta distribution with parameters α and β. Note that the I_i^T's are specified to have a common distribution, which will, to a certain extent, take into account the fact that the stated intentions are somewhat related.

Under the model (4.2) and (4.3), the Bayes estimate of I_i^T is given by

$$\hat{I}_i^T = \left(\frac{\alpha + \beta}{\alpha + \beta + 1}\right)\left(\frac{\alpha}{\alpha + \beta}\right) + \left(1 - \frac{\alpha + \beta}{\alpha + \beta + 1}\right) I_i^S, \tag{4.4}$$

where here and hereafter, I_i^S will be taken to be on the 0–1 scale. The marginal distribution of I_i^S (unconditional on I_i^T) is the negative hypergeometric distribution, sometimes called the beta-binomial. The exact form is not important here, because we will use only the facts that unconditionally,

$$E(I_i^S) = \frac{\alpha}{\alpha + \beta}$$

$$\text{var}(I_i^S) = \frac{1}{10}\left(\frac{\alpha}{\alpha + \beta}\right)\left(1 - \frac{\alpha}{\alpha + \beta}\right)\left(\frac{\alpha + \beta + 10}{\alpha + \beta + 1}\right). \tag{4.5}$$

(See Kendall and Stuart, 1977, Vol. 1, for more information on the beta-binomial distribution.)

By using (4.5) and the method of moments, α and β can be estimated. From the full data set in Table 2, we have $\bar{I}^S = .172$ with estimated variance = .091. Equating these to the expressions in (4.5) and solving for α and β, we get $\hat{\alpha} = .25$ and $\hat{\beta} = .43$. These yield the empirical Bayes estimate

$$\hat{I}_i^T = (.405)(.172) + (.595) I_i^S, \tag{4.6}$$

Table 3. Consumer Intent Estimates and Parameters

Intent group	Observed intent	True intent	Empirical Bayes estimate
0	0	.07	.07
.1–.3	.19	.19	.18
.4–.6	.51	.41	.37
.7–.9	.79	.48	.54
1	1	.53	.67
MSE	.729		.055

NOTE: MSE scaled by $\hat{\sigma}^2 = .091$.

which can be seen, once again, to be a weighted average of the grand mean (.172) and the individual intention.

The 447 people in the sample were contacted after the time period to learn whether or not a car had been purchased. Thus the parameter values are known. These values, together with the usual estimates (observed intent) and empirical Bayes estimates, are given in Table 3.

As expected, the empirical Bayes estimates are far superior to the observed intent, yielding a 93% improvement in MSE. Notice that the parameter values are much closer together than the observed intent, the phenomenon anticipated by the empirical Bayes estimates. In fact, the regression toward the mean is even more pronounced than that predicted by the empirical Bayes estimates.

Table 3 shows that the empirical Bayes estimates perform remarkably well, but seen in another light, their performance is startling. From (4.4) and (4.6), it can be seen that we are using estimates of I_i^T that are linear functions of I_i^S. Since we now have the parameter values, we can see what the best linear predictor is (in practice, this can never be done). A linear regression of the true intent on the stated intent yields the line $.10 + .47 I_i^S$ as the best possible linear predictor. Compare this to the empirical Bayes line $.07 + .595X$, and it can be seen that the empirical Bayes line is incredibly close to the best possible (but always unattainable) line. Imagine doing a regression of y on x without any y values! Figure 4 illustrates this.

Finally, the empirical Bayes method can tell us something

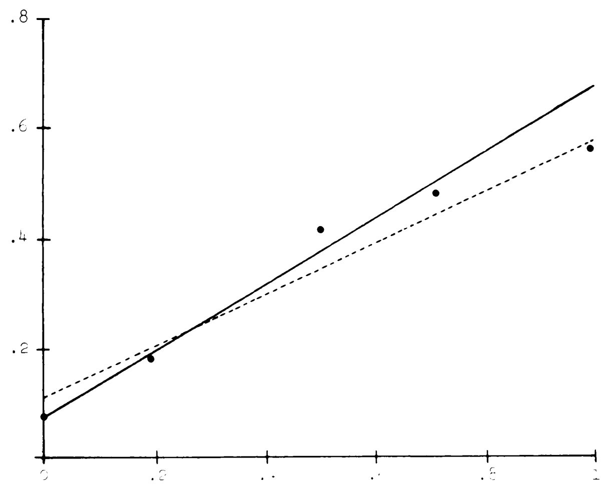

Figure 4. Comparison of the Empirical Bayes Estimate Line (——) With the Best Possible Linear Estimate Line (- - -). Values plotted are the weighted averages of the observed intent from Table 2.

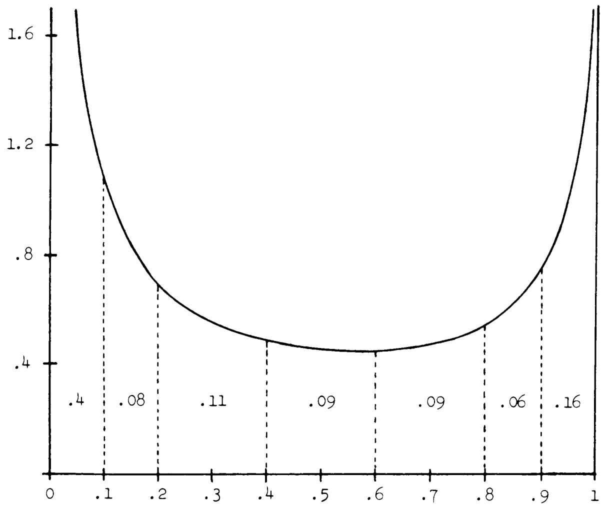

Figure 5. The Beta Density Function, α = .25, β = .43. The numbers under the curve represent the probability content of the indicated regions.

about the prior distribution, and such information can be useful, particularly if future studies are to be done. Recall that our estimates of α and β, the prior parameters, were .25 and .43, respectively. Figure 5 is a graph of the beta distribution with these parameter values. As one can see, the greatest concentration of mass is near the ends of the intervals, with the distribution being fairly flat in the middle. Since the beta distribution can have virtually any shape (from *U*-shape to bell-shape, symmetric or asymmetric), it is interesting that the empirical prior is an asymmetric *U*-shaped distribution. Since the empirical Bayes estimator produced such good estimates, it is reasonable to infer that this U-shaped prior is a reasonable approximation to the true prior distribution. Thus one would expect a population's true intentions to be clustered near 0 or 1, with a small portion (approximately 30%) uniformly distributed between .2 and .8.

[Received June 1984. Revised October 1984.]

REFERENCES

Berger, J. O. (1982), "Bayesian Robustness and the Stein Effect," *Journal of the American Statistical Association*, 77, 358–368.

Carter, G., and Rolph, J. (1974), "Empirical Bayes Methods Applied to Estimating Fire Alarm Probabilities," *Journal of the American Statistical Association*, 69, 880–885.

Efron, B., and Morris, C. (1972), "Limiting the Risk of Bayes and Empirical Bayes Estimators—Part II: The Empirical Bayes Case," *Journal of the American Statistical Association*, 67, 130–139.

——— (1973), "Stein's Estimation Rule and Its Competitors—An Empirical Bayes Approach," *Journal of the American Statistical Association*, 68, 117–130.

——— (1975), "Data Analysis Using Stein's Estimator and Its Generalizations," *Journal of the American Statistical Association*, 70, 311–319.

——— (1977), "Stein's Paradox in Statistics," *Scientific American*, 236 (5), 119–127.

Fay, R. E. III, and Herriot, R. A. (1979), "Estimates of Income for Small Places: An Application of James–Stein Procedures to Census Data," *Journal of the American Statistical Association*, 74, 269–277.

Hoadley, B. (1981), "Quality Management Plan (QMP)," *Bell System Technical Journal*, 60, 215–273.

Juster, F. T. (1966), "Consumer Buying Intentions and Purchase Probability: An Experiment in Survey Design," *Journal of the American Statistical Association*, 61, 658–696.

Kendall, M., and Stuart, A. (1977), *The Advanced Theory of Statistics* (Vol. 1; 4th ed.), New York: Macmillan.

Maritz, J. S. (1970), *Empirical Bayes Methods*, London: Methuen.

Morris, C. (1983), "Parametric Empirical Bayes Inference: Theory and Applications" (with discussion), *Journal of the American Statistical Association*, 78, 47–65.

Morrison, D. G. (1979), "Purchase Intentions and Purchase Behavior," *Journal of Marketing*, 43, 65–74.

Robbins, H. (1955), "An Empirical Bayes Approach to Statistics," *Proceedings of the Third Berkeley Symposium Mathematical Statistics and Probability* 1, Berkeley: University of California Press, pp. 157–164.

Rubin, D. (1981), "Using Empirical Bayes Techniques in the Law School Validity Studies," *Journal of the American Statistical Association*, 75, 801–827.

Stein, C. (1981), "Estimation of the Mean of a Multivariate Normal Distribution, *Annals of Statistics*, 9, 1135–1151.

Sutherland, M., Holland, P. W., and Fienberg, S. E. (1975), "Combining Bayes and Frequency Approaches to Estimate a Multinomial Parameter," in *Studies in Bayesian Econometrics and Statistics*, eds. S. E. Fienberg and A. Zellner, Amsterdam: North-Holland, pp. 585–617.