The Annals of Statistics
1983, Vol. 11, No. 3, 713–723

JERZY NEYMAN MEMORIAL LECTURE

SOME THOUGHTS ON EMPIRICAL BAYES ESTIMATION[1]

BY HERBERT ROBBINS

Columbia University

Examples are given to illustrate the "linear" and "general" empirical Bayes approaches to estimation. A final example concerns testing the null hypothesis that a treatment has had no effect.

1. The component problem. Let (θ, x) be a random vector such that (a) θ has a distribution function G, and (b) conditionally on θ, x has a probability density function $f(x \mid \theta)$ of known form with respect to a σ-finite measure m. We want to estimate θ by some function $t = t(x)$. The mean squared error of estimate with regard to the random variation of both θ and x is

$$E(t - \theta)^2 = \int\int (t(x) - \theta)^2 f(x \mid \theta)\, dm(x)\, dG(\theta). \tag{1}$$

For given G we can try to minimize (1) within any desired class of functions t. The minimum of (1) within the class of *linear* functions $A + Bx$ is attained for

$$\tilde{t}(x) = E\theta + \frac{\mathrm{Cov}(\theta, x)}{\mathrm{Var}\, x}(x - Ex), \tag{2}$$

and is equal to

$$E(\tilde{t} - \theta)^2 = \mathrm{Var}\, \theta - \frac{\mathrm{Cov}^2(\theta, x)}{\mathrm{Var}\, x}. \tag{3}$$

The minimum of (1) within the class of *all* Borel functions is attained for

$$t^*(x) = E(\theta \mid x) = \frac{\int \theta f(x \mid \theta)\, dG(\theta)}{\int f(x \mid \theta)\, dG(\theta)}, \tag{4}$$

and is equal to

$$E(t^* - \theta)^2 = E\, \mathrm{Var}(\theta \mid x). \tag{5}$$

If, as we shall assume, G is unknown to the statistician, then neither $\tilde{t}$ nor t^* can be used directly to estimate θ.

2. The empirical Bayes approach. Suppose that we are dealing with a large number N of independent versions of the component problem: (θ_i, x_i) for $i = 1, \cdots, N$ are i.i.d. random vectors such that the θ_i have distribution function G, while the probability

Received October 1982; revised February 1983.

[1] Research supported by Department of Energy, National Science Foundation, and National Institutes of Health.

AMS 1970 *subject classifications.* Primary 62C12; secondary 62C25.

Key words and phrases. Empirical Bayes, prediction, testing hypotheses.

density function of x_i conditionally on θ_i is $f(x \mid \theta_i)$. By observing $x_1, \cdots, x_N$, which are i.i.d. random variables with marginal p.d.f.

$$f(x) = \int f(x \mid \theta)\, dG(\theta), \tag{6}$$

we can try to gather enough information about G to approximate (2), a modest aim, or even (4), a more grandiose one. It is not obvious how this is to be done, nor whether we should try to approximate (2), (4), or perhaps something else.

One reason that we might be content with approximating (2) rather than (4) is that (2) involves only the constants Ex, $\operatorname{Var} x$, $E\theta$, and $\operatorname{Cov}(\theta, x)$, which for some families $f(x \mid \theta)$ are easy to estimate from $x_1, \cdots, x_N$. In addition, (2) often provides a minimax value of (1) within the class of all G with given $E\theta$ and $\operatorname{Var} \theta$. We may even believe that G belongs, or is close to belonging, to some special parametric family of distribution functions for which (4) is in fact a linear function of x, so that (2) and (4) are equal.

We refer to methods that seek to approximate (2) as *linear empirical Bayes* (l.e.B.), and to those that seek to approximate (4) as *general empirical Bayes* (g.e.B.). For some families $f(x \mid \theta)$ it might be desirable to replace the class of linear functions by the class of functions of the form $(x + A)/(x + B)$, or by some other class. We would then try to approximate whatever function in that class minimizes (1), referring to this as a *restricted empirical Bayes* (r.e.B.) method. For $N \to \infty$ the g.e.B. approach, when feasible, would seem to be better than any restricted approach, such as the linear, but for moderate N a suitable restricted approach with rapid convergence as $N \to \infty$ may be better than a general one that converges slowly.

If the θ_i are regarded not as random variables but simply as N arbitrary unknown constants, and if the loss function is the sum of the N squared errors of estimation, then the problem of simultaneously estimating all the θ_i is called a *compound* estimation problem. The same methods that apply to the e.B. problem will also apply to the compound problem, as indicated in Robbins (1951). Stein's (1956) inadmissibility result and the James-Stein (1961) estimator belong to the compound theory, but were put into an e.B. context in a series of papers by Efron and Morris; e.g., (1973). Except in Section 7 we confine ourselves in what follows to the e.B. theory, starting with l.e.B.

3. Linear empirical Bayes estimation.

We shall assume for simplicity that $f(x \mid \theta)$ is such that for all θ

$$E(x \mid \theta) = \theta. \tag{7}$$

Then

$$Ex = E\theta, \quad \operatorname{Cov}(\theta, x) = \operatorname{Var} \theta, \quad E(x - \theta)^2 = \operatorname{Var} x - \operatorname{Var} \theta = E \operatorname{Var}(x \mid \theta), \tag{8}$$

and (2) and (3) can be written as

$$\tilde{t}(x) = Ex + \frac{\operatorname{Var} \theta}{\operatorname{Var} x}(x - Ex), \quad E(\tilde{t} - \theta)^2 = \frac{\operatorname{Var} \theta \cdot E \operatorname{Var}(x \mid \theta)}{\operatorname{Var} \theta + E \operatorname{Var}(x \mid \theta)}. \tag{9}$$

We shall also assume, in addition to (7), that for some known constants a, b, c

$$\operatorname{Var}(x \mid \theta) = a + b\theta + c\theta^2. \tag{10}$$

Then

$$\operatorname{Var} x = E \operatorname{Var}(x \mid \theta) + \operatorname{Var} E(x \mid \theta) = a + bE\theta + cE^2\theta + (c + 1)\operatorname{Var} \theta, \tag{11}$$

so that

$$\operatorname{Var} \theta = \frac{\operatorname{Var} x - (a + bEx + cE^2x)}{c + 1}.$$

This suggests as an approximation to $\tilde{t}(x)$ the function

$$\tilde{\theta}(x) = \bar{x} + \left[1 - \frac{cs^2 + a + b\bar{x} + c\bar{x}^2}{(c+1)s^2}\right](x - \bar{x}), \tag{12}$$

where Ex and Var x in the first equation of (9) are conventionally estimated by

$$\bar{x} = \frac{1}{N}\textstyle\sum_1^N x_i \quad \text{and} \quad s^2 = \frac{1}{N-1}\textstyle\sum_1^N (x_i - \bar{x})^2. \tag{13}$$

(Of course, other estimators of Ex and Var x could be used. Moreover, the term in square brackets in (12) is an estimate of Var θ/Var x, which by (8) lies beween 0 and 1, so that we should truncate this term accordingly if sampling fluctuations of $\bar{x}$ and s^2 so require. We shall not bother to indicate this notationally.)

We call

$$\tilde{\theta}_i = \tilde{\theta}(x_i) = \bar{x} + \left[1 - \frac{cs^2 + a + b\bar{x} + c\bar{x}^2}{(c+1)s^2}\right](x_i - \bar{x}) \tag{14}$$

an l.e.B. estimator of θ_i, and under some mild restrictions on the nature of G we hope that with reasonable rapidity as $N \to \infty$

$$\begin{aligned} E(\tilde{\theta}_i - \theta_i)^2 &= \frac{1}{N}\textstyle\sum_1^N E(\tilde{\theta}_i - \theta_i)^2 \to E(\tilde{t} - \theta)^2 \\ &= \frac{\operatorname{Var}\theta \cdot E\operatorname{Var}(x\mid\theta)}{\operatorname{Var}\theta + E\operatorname{Var}(x\mid\theta)} = \frac{\operatorname{Var}\theta}{\operatorname{Var}x}E(x-\theta)^2, \end{aligned} \tag{15}$$

where for large N

$$\frac{\operatorname{Var}\theta}{\operatorname{Var}x} = \frac{\operatorname{Var}\theta}{a + bE\theta + cE^2\theta + (c+1)\operatorname{Var}\theta} \cong 1 - \frac{cs^2 + a + b\bar{x} + c\bar{x}^2}{(c+1)s^2}. \tag{16}$$

When the x_i are such that the last quantity is appreciably less than 1 it will seem advantageous to estimate θ_i by $\tilde{\theta}_i$ instead of by x_i, even if a decision to do so was not made in advance.

The most familiar case of (14) is that where for a known $\sigma > 0$, and for m = Lebesgue measure,

$$f(x\mid\theta) = \phi\left(\frac{x-\theta}{\sigma}\right)\Big/\sigma, \tag{17}$$

where ϕ is the p.d.f. of some random variable with mean 0 and variance 1, such as the standard normal. Then

$$E(x\mid\theta) = \theta, \quad \operatorname{Var}(x\mid\theta) = \sigma^2, \tag{18}$$

so that $a = \sigma^2$, $b = c = 0$ in (10), and (14) becomes

$$\tilde{\theta}_i = \bar{x} + \left[1 - \frac{\sigma^2}{s^2}\right](x_i - \bar{x}), \tag{19}$$

with

$$\frac{\operatorname{Var}\theta}{\operatorname{Var}x} = \frac{\operatorname{Var}\theta}{\sigma^2 + \operatorname{Var}\theta} \cong 1 - \frac{\sigma^2}{s^2}. \tag{20}$$

The estimator (19) is a variant of the James-Stein (1961) estimator of a multivariate normal mean referred to at the end of Section 2 above.

Before considering other examples, we shall generalize slightly the assumptions of Section 2 on the vectors (θ_i, x_i), which we now assume to be independent but not necessarily identically distributed, and such that

(21a) the θ_i have the same mean $E\theta$ and the same variance Var θ for all i

(21b) $E(x_i \mid \theta_i) = \theta_i$,

(21c) for some known constants a_i, b_i, c_i,

$$\operatorname{Var}(x_i \mid \theta_i) = a_i + b_i\theta_i + c_i\theta_i^2.$$

Then with s^2 defined by (13),

$$Ex_i = E\bar{x} = E\theta, \quad Es^2 = 1/N \textstyle\sum_1^N \operatorname{Var} x_i = \bar{a} + \bar{b}E\theta + \bar{c}E^2\theta + (\bar{c} + 1)\operatorname{Var}\theta, \tag{22}$$

where $\bar{a} = \sum_1^N a_i/N$, etc. If, without regard to efficiency, we simply estimate

$$E\theta \quad \text{by} \quad \bar{x},$$

$$\operatorname{Var}\theta \quad \text{by} \quad \frac{s^2 - (\bar{a} + \bar{b}\bar{x} + \bar{c}\bar{x}^2)}{\bar{c} + 1},$$

and

$$\operatorname{Var} x_i = E \operatorname{Var}(x_i \mid \theta_i) + \operatorname{Var} E(x_i \mid \theta_i) = a_i + b_iE\theta + c_iE^2\theta + (c_i + 1)\operatorname{Var}\theta$$

$$\text{by} \quad a_i + b_i\bar{x} + c_i\bar{x}^2 + \frac{(c_i + 1)}{(\bar{c} + 1)}\{s^2 - (\bar{a} + \bar{b}\bar{x} + \bar{c}\bar{x}^2)\},$$

then we can approximate

$$\tilde{t}(x_i) = E\theta + \frac{\operatorname{Var}\theta}{\operatorname{Var} x_i}(x_i - E\theta) \tag{23}$$

by

$$\tilde{\theta}_i = \bar{x} + \left[1 - \frac{(\bar{c} + 1)(a_i + b_i\bar{x} + c_i\bar{x}^2) + c_i\{s^2 - (\bar{a} + \bar{b}\bar{x} + \bar{c}\bar{x}^2)\}}{(\bar{c} + 1)(a_i + b_i\bar{x} + c_i\bar{x}^2) + (c_i + 1)\{s^2 - (\bar{a} + \bar{b}\bar{x} + \bar{c}\bar{x}^2)\}}\right](x_i - \bar{x}). \tag{24}$$

We continue to call (24) an l.e.B. estimate of θ_i under the assumptions (21). When $a_i = a$, $b_i = b$, $c_i = c$, (24) reduces to (14).

4. Examples of l.e.B. estimation.

EXAMPLE 1. For $i = 1, \cdots, N$ let y_{ij} for $j = 1, \cdots, n_i$ be independent, conditionally on $\theta_1, \cdots, \theta_N$, with mean θ_i and known variance σ_i^2, and let

$$x_i = \frac{1}{n_i}\textstyle\sum_{j=1}^{n_i} y_{ij} = i\text{th sample mean.} \tag{25}$$

Then

$$E(x_i \mid \theta_i) = \theta_i, \quad \operatorname{Var}(x_i \mid \theta_i) = \frac{\sigma_i^2}{n_i} = a_i. \tag{26}$$

Assuming that the θ_i are independent with a common mean and variance, (24) yields the l.e.B. estimator

$$\tilde{\theta}_i = \bar{x} + \left[1 - \frac{a_i}{(a_i - \bar{a}) + s^2}\right](x_i - \bar{x}). \tag{27}$$

Of course, it is inefficient to estimate $E\theta$ by $\bar{x}$ instead of by the weighted average

$$\textstyle\sum_1^N (n_ix_i/\sigma_i^2)/\sum_1^N (n_i/\sigma_i^2), \tag{28}$$

and in practice we would use (27) only when the a_i are nearly equal. Many other modifications of (27) could be made to improve its performance for moderate N. We shall not pursue these questions here, although they are important in the practical application of l.e.B. methods.

EXAMPLE 2. As in Example 1 but with the σ_i^2 unknown. Defining

$$s_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (y_{ij} - x_i)^2 = i\text{th sample variance,} \tag{29}$$

and assuming that the n_i are large, we could replace the $a_i = \sigma_i^2/n_i$ in (27) by $a_i = s_i^2/n_i$ and still regard (27) as an l.e.B. estimate of θ_i. However, if the y_{ij} are *normal*, then setting $r_i = (n_i - 1)/2$ we have

$$E(s_i^2 \mid \sigma_i^2) = \sigma_i^2, \quad \mathrm{Var}(s_i^2 \mid \sigma_i^2) = \sigma_i^4/r_i, \tag{30}$$

so that (21b) holds with θ_i replaced by σ_i^2 and x_i by s_i^2, and (21c) holds with $a_i = b_i = 0$, $c_i = 1/r_i$. If we regard the σ_i^2 as independent random variables with a common mean and variance, we obtain from (24) the l.e.B. estimates (cf. Robbins (1982), formula (49))

$$\tilde{\sigma}_i^2 = q^2 + \left[1 - \frac{R^4 + q^4}{R^4 + q^4 + r_i(R^4 - \bar{c}q^4)}\right](s_i^2 - q^2), \tag{31}$$

where by definition

$$q^2 = \frac{1}{N}\sum_1^N s_i^2, \quad R^4 = \frac{1}{N-1}\sum_1^N (s_i^2 - q^2)^2, \quad \bar{c} = \frac{1}{N}\sum_1^N \frac{1}{r_i}. \tag{32}$$

For large N we could then use (27) with $a_i = \tilde{\sigma}_i^2/n_i$ to obtain "doubly" l.e.B. estimates of the population means θ_i. The assumption of normality for the y_{ij} can be dropped by a suitable generalization of (31).

Another l.e.B. approach that dispenses with normality to the simultaneous estimation of many population means $\theta_1, \cdots, \theta_N$ is the following. Let $F_1, \cdots, F_N$ be i.i.d. *random distribution functions*, and for given F_i ($i = 1, \cdots, N$) let the y_{ij} ($j = 1, \cdots, n_i$) be distributed according to F_i. Define x_i by (25), s_i^2 by (29), $\bar{x}$ and s^2 by (13), and let

$$\theta_i = \int t\, dF_i(t) = \text{random mean of population } i = E(y_{ij} \mid F_i),$$

$$\sigma_i^2 = \int (t - \theta_i)^2\, dF_i(t) = \text{random variance of population } i = \mathrm{Var}(y_{ij} \mid F_i).$$

Then $\theta_1, \cdots, \theta_N$ are i.i.d. random variables, as are $\sigma_1^2, \cdots, \sigma_N^2$. We observe that

$$E(x_i \mid F_i) = \theta_i, \quad Ex_i = E\theta_i = E\theta = E\bar{x}, \quad \mathrm{Cov}(\theta_i, x_i) = \mathrm{Var}\,\theta,$$

$$\mathrm{Var}\,x_i = E\,\mathrm{Var}(x_i \mid F_i) + \mathrm{Var}\,E(x_i \mid F_i) = E\,\mathrm{Var}(x_i \mid F_i) + \mathrm{Var}\,\theta$$

$$Es^2 = (1/N)\sum_1^N \mathrm{Var}\,x_i = (1/N)\sum_1^N E\,\mathrm{Var}(x_i \mid F_i) + \mathrm{Var}\,\theta,$$

so that (23) can be written as

$$\tilde{t}(x_i) = E\theta + \left[\frac{Es^2 - (1/N)\sum_1^N E\,\mathrm{Var}(x_i \mid F_i)}{Es^2 + E\,\mathrm{Var}(x_i \mid F_i) - (1/N)\sum_1^N E\,\mathrm{Var}(x_i \mid F_i)}\right](x_i - E\theta). \tag{33}$$

But, defining q^2 by (32), we have

$$\mathrm{Var}(x_i \mid F_i) = \frac{\sigma_i^2}{n_i}, \quad E(s_i^2 \mid F_i) = \sigma_i^2, \quad Es_i^2 = E\sigma_i^2 = E\sigma^2 = Eq^2,$$

so that for large N,

$$E\operatorname{Var}(x_i \mid F_i) = \frac{E\sigma^2}{n_i} \cong \frac{q^2}{n_i}, \quad Es^2 \cong s^2, \quad E\theta \cong \bar{x}.$$

Hence, we can approximate $\tilde{t}(x_i)$ by (27) with

$$a_i = \frac{q^2}{n_i} = \frac{1}{n_i}\left(\frac{1}{N}\sum_1^N s_i^2\right), \tag{34}$$

in contrast to $a_i = \sigma_i^2/n_i$ known, $a_i = s_i^2/n$, or $a_i = \tilde{\sigma}_i^2/n_i$ with $\tilde{\sigma}_i^2$ given by (31) when the F_i are normal.

EXAMPLE 3. Let y_i denote the number of successes in n_i Bernoulli trials with success probability p_i, and let $x_i = y_i/n_i$. Then

$$E(x_i \mid p_i) = p_i, \quad \operatorname{Var}(x_i \mid p_i) = p_i(1 - p_i)/n_i,$$

so that in (21c) $a_i = 0$, $b_i = -c_i = 1/n_i$, and the l.e.B. estimate (24) of $\theta_i = p_i$ becomes

$$\tilde{p}_i = \bar{x} + \left[1 - \frac{\bar{x}(1-\bar{x}) - s^2}{\bar{x}(1-\bar{x}) - s^2 + n_i\{s^2 - \bar{b}\bar{x}(1-\bar{x})\}}\right](x_i - \bar{x}).$$

This estimate was introduced by Copas (1972) with a slightly different motivation.

EXAMPLE 4. Returning to (14), let x denote the number of successes before the first failure in a sequence of Bernoulli trials with success probability $p = 1 - q$. Then

$$f(x \mid p) = qp^x \quad (x = 0, 1, \cdots) \tag{35}$$

and

$$E(x \mid p) = p/q, \quad \operatorname{Var}(x \mid p) = p/q^2. \tag{36}$$

To satisfy (7) and (10) we introduce $\theta = p/q$ as the parameter to be estimated, so that

$$E(x \mid \theta) = \theta, \quad \operatorname{Var}(x \mid \theta) = \theta(1 + \theta), \tag{37}$$

with $a = 0$, $b = c = 1$ in (10). Then (14) becomes

$$\tilde{\theta}_i = \bar{x} + \left[1 - \frac{s^2 + \bar{x}(1 + \bar{x})}{2s^2}\right](x_i - \bar{x}), \tag{38}$$

which by (15) and (16) improves on x_i for large N by the factor

$$\frac{\operatorname{Var}\theta}{\operatorname{Var}x} = \frac{\operatorname{Var}\theta}{E\theta(1 + E\theta) + 2\operatorname{Var}\theta} \cong \left\{1 - \frac{\bar{x}(1 + \bar{x})}{s^2}\right\} \Big/ 2. \tag{39}$$

5. G.e.B. estimation; an example. The problem of estimation for the examples $f(x \mid \theta)$ of the preceding section, along with many others, can also be considered from the g.e.B. aspect of approximating (4) instead of (2), when the component problems are identical and N is large. We shall treat here only Example 4, the geometric distribution with parameter $\theta = p/q$, for which (6) becomes

$$f(x) = \int_0^1 qp^x \, dG(p) \tag{40}$$

and the unrestrictedly best estimator of θ is

$$t^*(x) = E(\theta \mid x) = \int_0^1 p^{x+1} \, dG(p)/f(x). \tag{41}$$

Since

$$p^{x+1} = \textstyle\sum_{i=0}^{\infty} qp^{x+i+1}, \tag{42}$$

we can write $t^*(x)$ in terms of $f(x)$ as

$$t^*(x) = [f(x+1) + f(x+2) + \cdots]/f(x). \tag{43}$$

This suggests defining a g.e.B. estimate of θ_i by

$$\theta_i^* = \frac{\text{number of terms } x_1, \cdots, x_N \text{ that are greater than } x_i}{\text{number of terms } x_1, \cdots, x_N \text{ that are equal to } x_i}. \tag{44}$$

We remark that, contrary to the situation in Example 4 of the preceding section, it is easy to obtain a g.e.B. estimator of p_i itself rather than of $\theta_i = p_i/q_i$. Since

$$E(p \mid x) = \int_0^1 qp^{x+1}\, dG(p)/f(x) = f(x+1)/f(x), \tag{45}$$

it is natural to define a g.e.B. estimate of p_i by

$$p_i^* = \frac{\text{number of terms } x_1, \cdots, x_N \text{ that are equal to } (x_i + 1)}{\text{number of terms } x_1, \cdots, x_N \text{ that are equal to } x_i}. \tag{46}$$

We have in (38) and (44) two quite different e.B. estimators of $\theta_i = p_i/q_i$ for the case (35). But when G belongs to the Beta family, with

$$G'(p) = g(p) = \frac{1}{B(a, b)} p^{a-1} q^{b-1} \quad (a, b > 0), \tag{47}$$

then (2) and (4) are equal, so that (38) and (44) are approximations to the same thing. In fact, when (47) holds (40) becomes

$$f(x) = \frac{B(x+a, b+1)}{B(a, b)}, \tag{48}$$

and hence by (41)

$$t^*(x) = \frac{B(x+1+a, b)}{B(x+a, b+1)} = \frac{x+a}{b} = \tilde{t}(x). \tag{49}$$

From (7), (8), and (11) it follows that

$$Ex + \frac{\operatorname{Var} x - Ex(1+Ex)}{2 \operatorname{Var} x}(x - Ex) = \frac{x+a}{b}, \tag{50}$$

so that

$$\begin{aligned} b &= \frac{2 \operatorname{Var} x}{\operatorname{Var} x - Ex(1+Ex)} \cong \frac{2s^2}{s^2 - \bar{x}(1+\bar{x})}, \\ a &= Ex \cdot \left[\frac{\operatorname{Var} x + Ex(1+Ex)}{\operatorname{Var} x - Ex(1+Ex)}\right] \cong \frac{\bar{x}\{s^2 + \bar{x}(1+\bar{x})\}}{s^2 - \bar{x}(1+\bar{x})}, \\ \frac{x+a}{b} &\cong \bar{x} + \left[1 - \frac{s^2 + \bar{x}(1+\bar{x})}{2s^2}\right](x - \bar{x}) = (38) \text{ with } x_i = x. \end{aligned} \tag{51}$$

Thus, for the family of Beta priors the l.e.B. estimator (38) of θ_i is an estimate by the method of moments of $t^*(x_i)$. (The g.e.B. estimator (44) is a more generally valid estimate of $t^*(x_i)$, but presumably converges more slowly to (41) as $N \to \infty$.) We could, of course, estimate a and b in (49) by maximum likelihood, finding for given $x_1, \cdots, x_N$ the maximum with respect to a, b of

$$\prod_{i=1}^{N} \frac{B(x_i + a,\, b + 1)}{B(a,\, b)},$$

but this is a dubious procedure if there is a possibility of a non-Beta prior. The virtue of (38) would seem to lie in the fact that for large N it is nearly equal to (2) for *any* G. (It is also true that

$$\tilde{t}(x) = E\theta + \frac{\operatorname{Var}\theta}{E\theta(1 + E\theta) + 2\operatorname{Var}\theta}(x - E\theta) \tag{52}$$

is minimax in the class of all priors with given $E\theta$ and Var θ, but no great importance attaches to this fact in itself.) For very large N, (44) would seem to be preferable to (38), unless there is some iron-clad reason to believe that G is Beta. A method of combining the best features of l.e.B. and g.e.B. is given in Robbins (1980), but finding a more or less optimal way of treating both moderate and large values of N will require considerable theoretical and computational study.

6. Prediction and testing. We have thus far considered only the usual e.B. problem of simultaneous estimation of many parameters. We consider now a problem (cf. Robbins, 1977) in which no parameters at all are to be estimated.

Let $\theta_1, \cdots, \theta_N$ be i.i.d. with unknown distribution function G, and conditionally on θ_i let x_i and y_i be independent random variables with same probability density function $f(\cdot \mid \theta_i)$. Thus, (x_i, y_i) for $i = 1, \cdots, N$ are i.i.d. random vectors with joint marginal p.d.f.

$$f(x, y) = \int f(x \mid \theta) f(y \mid \theta)\, dG(\theta). \tag{53}$$

We suppose that the x_i have been observed but not the y_i; the θ_i, in practice, will never be observed. Let A be some set of interest in the x-space, and let

$$S = \textstyle\sum_{x_i \in A} y_i \tag{54}$$

be *the sum of all the not-yet-observed y_i values for which the corresponding observed x_i values belong to A*. Problem: find a good function $g(x_1, \cdots, x_N)$ to use as a predictor of S.

For example, N typists are chosen at random from some population about which nothing is known, and each types a text until he or she makes an error. Let x_i denote the number of symbols typed correctly by typist i, and let ν denote the number of typists whose x_i values are $\geq a$, where a is some fixed positive integer. These ν typists are to be set to typing again, until each makes an error. Let S denote the total number of symbols that will be typed correctly by the ν typists in this second round. From the values $x_1, \cdots, x_N$, predict S. (It is assumed that for typist i, the typing of a correct or incorrect symbol forms a sequence of Bernoulli trials with constant probability p_i of success throughout both rounds, p_i being characteristic of typist i but varying with i in some unknown manner, since nothing is known about the population from which the N typists were randomly selected.)

Clearly, ν times the average of $x_1, \cdots, x_N$ will generally underestimate S, while the sum of the ν values of x_i that are $\geq a$ will generally overestimate S. What, then, is a reasonable predictor for S, and how accurate is it likely to be? We shall treat this as a g.e.B. problem.

For the density (35), (53) becomes

$$f(x, y) = \int_0^1 q^2 p^{x+y}\, dG(p), \tag{55}$$

a function of $x + y$ *only*. It follows that

(56) conditionally on the value of $z = x + y$,
the distribution of x is uniform on the set $\{0, 1, 2, \cdots, z\}$.

From now on we ignore the explicit form (55) of $f(x, y)$ and assume only (56), which is more general.

Let (x, y) be any random vector with $x, y = 0, 1, \cdots$ such that (56) holds, and let (x_i, y_i), $i = 1, \cdots, N$, be i.i.d. with the same distribution as (x, y). Let

$$S = \textstyle\sum_{x_i \geq a} y_i = \sum_1^N u(x_i) y_i, \tag{57}$$

where by definition

$$u(x) = \begin{array}{ll} 1 & \text{if} \quad x \geq a \\ 0 & \text{if} \quad x < a. \end{array} \tag{58}$$

By (56),

$$E[u(x) y \mid z] = \frac{1}{z + 1} \textstyle\sum_{x=a}^{z} (z - x) = \dfrac{1 + 2 + \cdots + (z - a)}{z + 1} \tag{59}$$

for $z \geq a$, and is 0 for $z < a$. Now define

$$v(x) = (z - a)^+ = \begin{array}{ll} x - a & \text{if} \quad x \geq a \\ 0 & \text{if} \quad x < a \end{array}, \tag{60}$$

and observe that

$$E[v(x) \mid z] = \frac{1}{z + 1} \textstyle\sum_{x=a}^{z} (x - a) = \dfrac{1 + 2 + \cdots + (z - a)}{z + 1} \tag{61}$$

for $z \geq a$, and is 0 for $z < a$. Since z was arbitrary, it follows from (59) and (61) that $E[u(x) y] = E[v(x)]$, and hence that *the statistic*

$$T = \textstyle\sum_1^N v(x_i) \tag{62}$$

has the same expected value as S. Thus,

$$S - T = \textstyle\sum_1^N (u(x_i) y_i - v(x_i)) \tag{63}$$

is the sum of N i.i.d. random variables with mean 0, so that as $N \rightarrow \infty$

$$\frac{S - T}{\sigma \sqrt{N}} \rightarrow N(0, 1) \quad \text{in distribution,} \tag{64}$$

where by definition

$$\sigma^2 = E[u(x) y - v(x)]^2 = E[u(x) \{y - v(x)\}^2], \tag{65}$$

and

$$\begin{aligned} E[u(x)\{y - v(x)\}^2 \mid z] &= E[u(x)\{z - x - (x - a)\}^2 \mid z] \\ &= E[u(x)\{z - 2x + a\}^2 \mid z] = \frac{1}{z + 1} \textstyle\sum_{x=a}^{Z} (z - 2x + a)^2 \\ &= \frac{1}{z + 1} \textstyle\sum_{i=0}^{k} (k - 2i)^2 \quad \text{where} \quad k = z - a \\ &= \frac{k(k + 1)(k + 2)}{3(z + 1)} \end{aligned} \tag{66}$$

for $z \geq a$, and is 0 for $z < a$.

We now define

$$w(x) = (x - a)^+(x - a + 1), \tag{67}$$

and observe that

$$E[w(x) \mid z] = \frac{1}{z+1}\textstyle\sum_{x=a}^{z} (x - a)(x - a + 1) = \frac{1}{z+1}\sum_{i=0}^{k} i(i + 1) = \frac{k(k + 1)(k + 2)}{3(z + 1)}$$

for $z \geq a$, and is 0 for $z < a$. Hence $Ew(x) = \sigma^2$.

It follows that as $N \to \infty$

$$(1/N) \textstyle\sum_1^N w(x_i) \to Ew(x) = \sigma^2 \quad \text{in probability,} \tag{68}$$

and hence from (64) that as $N \to \infty$

$$\frac{S - T}{\sqrt{\sum_1^n w(x_i)}} \to N(0, 1) \quad \text{in distribution,} \tag{69}$$

so that *for large N an approximately* 95% *prediction interval for S is given by*

$$\textstyle\sum_1^N (x_i - a)^+ \pm 1.96\sqrt{\sum_1^N (x_i - a)^+(x_i - a + 1)}. \tag{70}$$

This quantifies in the case of (56) the general phenomenon of "regression toward mediocrity."

If it should turn out that S does not lie in the interval (70), it would be reasonable to suspect that the ν typists with scores $x_i \geq a$ on the first round have changed their performance characteristics p_i on the second round. If some "treatment" has in fact been given to these typists between the first and second rounds, we can test the null hypothesis H_0 that the treatment has had no effect *even though no randomly chosen subset of the ν has been singled out as a control group*; cf. Robbins (1982).

7. Concluding remarks.

Neyman (1962) brought the compound and e.B. approaches to the attention of the general statistical public, and gave great impetus to the subsequent development of the theory. According to a recent note of Forcini (1982), the e.B. approach was anticipated by Gini in 1911.

We conclude with the following estimation theorem (cf. Tsui and Press, 1982), the proof of which is an interesting elementary exercise. Let $x_1, \cdots, x_N$ be independent Poisson random variables with respective means $\theta_1, \cdots, \theta_N$, and let $\bar{x} = \sum x_i/N$, $\bar{\theta} = \sum \theta_i/N$, $\alpha = \bar{\theta}/(1 + \bar{\theta})$. Then for any $N \geq 2$

$$E\left\{\frac{1}{N}\sum_{i=1}^{N} \frac{\left(\frac{\bar{x}x_i}{1 + \bar{x}} - \theta_i\right)^2}{\theta_i}\right\} < \alpha + \frac{1}{N}(1 - \alpha^2) < 1. \tag{71}$$

As stated, (71) belongs to the compound theory; it becomes e.B. if the θ_i, as well as the x_i, are random variables, possibly dependent, but with finite expectations. Then the same inequality holds, with α equal to $\mu/(1 + \mu)$, $\mu = \sum E\theta_i/N$.

I think that Neyman would have appreciated this result.

REFERENCES

Copas, J. B. (1972). Empirical Bayes methods and the repeated use of a standard. *Biometrika* **59** 349-360.

EFRON, B. and MORRIS, C. (1973). Stein's estimation rule and its competitors—an empirical Bayes approach. *J. Amer. Statist. Assoc.* **68** 117–130.

FORCINI, A. (1982). Gini's contributions to the theory of inference. *Int. Statist. Rev.* **50** 65–70.

JAMES, W. and STEIN, C. (1961). Estimation with quadratic loss. *Proc. Fourth Berkeley Symp. Math. Statist. Probab.* 361–379. Univ. of Calif. Press.

NEYMAN, J. (1962). Two breakthroughs in the theory of statistical decision making. *Rev. Intern. Statist. Inst.* **30** 11–27.

ROBBINS, H. (1951). Asymptotically subminimax solutions of compound statistical decision problems. *Proc. Second Berkeley Symp. Math Statist. Probab.* 131–148. Univ. of Calif. Press.

ROBBINS, H. (1956). An empirical Bayes approach to statistics. *Proc. Third Berkeley Symp. Math. Statist. Probab.* **1** 157-163. Univ. of Calif. Press.

ROBBINS, H. (1977). Prediction and estimation for the compound Poisson distribution. *Proc. Natl. Acad. Sci. U.S.A.* **74** 2670-2671. See also Robbins, H. (1979). Some estimation problems for the compound Poisson distribution. *Asymptotic Theory of Statistical Tests and Estimation* 251-257. Academic, New York.

ROBBINS, H. (1980). An empirical Bayes estimation problem. *Proc. Natl. Acad. Sci. USA* **77** 6988–6989.

ROBBINS, H. (1982). Estimating many variances. *Statistical Decision Theory and Related Topics* III, Vol. 2, 218-226. Academic, New York.

STEIN, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. *Proc. Third Berkeley Symp. Math. Statist. Probab.* **1** 197-206. Univ. of Calif. Press.

TSUI, K.-W. and PRESS, S. J. (1982). Simultaneous estimation of several Poisson parameters under k-normalized squared error loss. *Ann. Statist.* **10** 93–100.

DEPARTMENT OF STATISTICS
COLUMBIA UNIVERSITY
NEW YORK, NEW YORK 10027
and
APPLIED MATHEMATICS DEPARTMENT
BROOKHAVEN NATIONAL LABORATORY
UPTON, NEW YORK 11973