

SOME ADMISSIBLE EMPIRICAL BAYES PROCEDURES¹

BY GLEN MEEDEN

Iowa State University

1. Summary. This paper considers some empirical Bayes procedures which have been discussed by H. Robbins, E. Samuel, and M. V. Johns. These procedures are shown to be inadmissible relative to a class of priors and by using some of the results of Rolph [6] admissible procedures are found for two examples. For an introduction to the empirical Bayes approach see Robbins [5].

2. Introduction. Let X be a discrete random variable with a family of possible probability distributions indexed by $\lambda \in \Omega$, an interval of real numbers. When the parameter is λ , X has the specified probability function f_λ . We are faced with a specified statistical decision problem which is given by D , the space of possible decisions, and L , a nonnegative loss function. If δ is a decision function then its risk function is $r(\delta, \lambda) = \sum_x L(\delta(x), \lambda) f_\lambda(x)$.

In the Bayes approach the parameter λ is the realization of a random variable Λ , distributed according to some a priori distribution function G on Ω . The Bayes risk of a decision function δ relative to the a priori distribution G is $R(\delta, G) = \int_\Omega r(\delta, \lambda) dG(\lambda)$. Any decision function, δ_G , satisfying $R(\delta_G, G) = R(G) = \min_\delta R(\delta, G)$ is called a Bayes decision function relative to G .

In the empirical Bayes approach the Bayes problem just described occurs repeatedly and independently with the same unknown G throughout. At the time when the decision about λ_{n+1} is to be made we have observed $(X^{(n)}, X_{n+1})$ where $X^{(n)} = (X_1, \dots, X_n)$ (the values $\lambda_1, \lambda_2, \dots$ always remaining unknown). Therefore, we can use for the decision about λ_{n+1} a decision function

$$(1) \quad \delta(x^{(n)}, \cdot).$$

A sequence $\delta = \{\delta_n\}$ where each δ_n , $n = 1, 2, \dots$ is of form (1) is called an empirical decision procedure. For each n the risk is given by

$$\bar{R}(\delta_n, G) = \sum_{x^{(n)}} R(\delta_n(x^{(n)}, \cdot), G) f_G^n(x^{(n)})$$

where

$$f_G^n(x^{(n)}) = f_G(x_1) f_G(x_2) \cdots f_G(x_n),$$

$$f_G(x_i) = \int_\Omega f_\lambda(x_i) dG(\lambda) \quad \text{for } i = 1, \dots, n$$

and $\delta_n(x^{(n)})$ denotes the decision function $\delta_n(x^{(n)}, \cdot)$. If $\lim_{n \rightarrow \infty} \bar{R}(\delta_n, G) = R(G)$ we say that δ is asymptotically optimal (a.o.) relative to G or that it is empirical Bayes. (See Robbins [5].)

Received January 24, 1969; revised June 14, 1971.

¹ This research was supported in part by National Science Foundation Grant GP 7363.

Let δ_n be given by (1) and \mathcal{G} be a class of possible a priori distributions on Ω . δ_n is said to be inadmissible relative to \mathcal{G} if there exists another decision function, δ_n^* , of form (1) such that $\bar{R}(\delta_n^*, G) \leq \bar{R}(\delta_n, G)$ for all $G \in \mathcal{G}$, and with strict inequality for at least one G . If δ_n is not inadmissible relative to \mathcal{G} it is said to be admissible relative to \mathcal{G} . A sequence $\delta = \{\delta_n\}$ is admissible if each δ_n is admissible.

Samuel [7], Johns [1], and Robbins [4] have exhibited procedures which are a.o. relative to a large class of possible a priori distributions, say \mathcal{G} . It will be shown that these sequences are inadmissible relative to \mathcal{G} because each member of the sequences, with the possible exception of the first, is inadmissible relative to \mathcal{G} . It is not surprising that these sequences are inadmissible since they were developed for their asymptotic properties. However, other considerations being equal, we would prefer an admissible a.o. sequence to one that is not.

3. The inadmissibility of some sequences. The proof of inadmissibility is based on the following observation. Let δ_n (n fixed) be an element of some empirical decision procedure δ , then

$$(2) \quad \bar{R}(\delta_n, G) = \sum_{x^{(n)}} \int_{\Omega} r(\delta_n(x^{(n)}), \lambda) dG(\lambda) f_G^n(x^{(n)}),$$

where $\delta_n(x^{(n)})$ denotes the decision function $\delta_n(x^{(n)}, \cdot)$. If for some fixed vector $x_0^{(n)}$, $\delta_n(x_0^{(n)}, \cdot)$ is inadmissible then there exists a decision function δ' such $r(\delta', \lambda) \leq r(\delta_n(x_0^{(n)}), \lambda)$ for all $\lambda \in \Omega$ with strict inequality for at least one λ . If $r(\delta, \lambda)$ is a continuous function λ , then the previous inequality must be strict on some interval of λ 's, say I contained in Ω . If $\delta'(x^{(n)}, x_{n+1}) = \delta'(x_{n+1})$ when $x^{(n)} = x_0^{(n)}$ and equals $\delta(x^{(n)}, x_{n+1})$ otherwise, then $\bar{R}(\delta_n', G) < \bar{R}(\delta_n, G)$ for any G which puts positive probability on I and with $f_G^n(x_0^{(n)}) > 0$. Hence δ_n is inadmissible relative to any class \mathcal{G} which contains such an a priori distribution G .

EXAMPLE 1. Suppose X is geometric on the positive integers with $f_i(x) = (1 - \lambda)\lambda^{x-1}$ for $\lambda \in [0, 1) = \Omega$. Consider the testing problem $H_0: \lambda \leq \lambda^*$ against $H_1: \lambda > \lambda^*$. Let d_i be the decision that we accept H_i for $i = 0, 1$. A decision function δ is a function such that $0 \leq \delta(x) \leq 1$ for all x where $\delta(x)$ denotes the probability of deciding d_1 given that $X = x$ is observed. The loss function is $L(d_0, \lambda) = \lambda - \lambda^*$ or 0 as $\lambda > \lambda^*$ or $\lambda \leq \lambda^*$ and $L(d_1, \lambda) = \lambda^* - \lambda$ or 0 as $\lambda < \lambda^*$ or $\lambda \geq \lambda^*$. By Theorem 3 on page 72 of Lehmann [3] and the completeness of the family of distributions, any test not given by $\delta(x) = 1$ or 0 as $x > c$ or $x < c$ for some c is inadmissible.

Now we assume that λ is a random variable and we are faced with the empirical Bayes situation. Samuel [7] defines the sequence (δ_n) as follows

$$(3) \quad \begin{aligned} \delta_n(x^{(n)}, x_{n+1}) &= 1 && \text{for } f_n(x_{n+1} + 1)/f_n(x_{n+1}) > \lambda^*, \\ &= 0 && \text{otherwise,} \end{aligned}$$

where $f_n(x_{n+1}) = f_n(x^{(n)}, x_{n+1}) =$ (number of indices $i, i = 1, 2, \dots, n$ for which $x_i = x_{n+1}$) and shows that (δ_n) is a.o. relative to the class of all a priori distributions on $[0, 1)$. Denote this class by \mathcal{S} . Let $n \geq 2$ be a fixed integer, then there exists a vector $x_0^{(n)}$ such that $f_n(1) = 1, f_n(2) = 1,$ and $f_n(3) = 0$ and therefore $\delta_n(x_0^{(n)}, \cdot)$ is inadmissible and by the remarks following (2) δ_n is inadmissible relative to \mathcal{S} .

Samuel also constructs a.o. sequences for the corresponding Poisson and negative binomial testing problems. The preceding argument proves inadmissibility in these cases too.

EXAMPLE 2. Let X be a Poisson (λ) variable with $\lambda \in (0, +\infty) = \Omega$. For estimating λ with squared error as the loss function it follows from Karlin and Rubin [2] and the completeness of the Poisson family that any estimator which is not a non-decreasing function is inadmissible.

Assume now that we are faced with the empirical Bayes situation. Robbins [5] defines the sequence (δ_n) as follows:

$$\delta_n(x^{(n)}, x_{n+1}) = (x_{n+1} + 1)f_n(x_{n+1} + 1)/(1 + f_n(x_{n+1})),$$

where $f_n(x_{n+1})$ is as in Example 1. Johns [1] shows that (δ_n) is a.o. relative to every member of the class $\mathcal{S} = \{G: \int_0^\infty \lambda^2 dG(\lambda) < \infty\}$. Let n be a fixed positive integer. As in Example 1 it is easy to find a vector $x_0^{(n)}$ such that $\delta_n(x_0^{(n)}, \cdot)$ is inadmissible and so δ_n is inadmissible relative to \mathcal{S} .

If in Example 1 we had wished to estimate λ with squared error loss Robbins [5] has exhibited a sequence of empirical estimators which is a.o. relative to the class of all a priori distributions on $[0, 1)$. The preceding argument also shows the inadmissibility of this sequence.

4. Admissible empirical Bayes procedures. In this section we will give two examples of admissible a.o. procedures. These procedures are constructed by treating the choice of δ_n as a Bayes decision problem with risk $\bar{R}(\delta_n, G)$, G playing the role of the unknown parameter. A prior distribution for G over the family \mathcal{S} is introduced and we choose a δ_n^* which minimizes the "posterior risk", that is, minimizes the average of \bar{R} with respect to the posterior distribution over \mathcal{S} . The Bayes nature of the solution can be expected to assure admissibility. If, as usually happens, the posterior distribution converges to the degenerate distribution at G_0 , the true prior distribution for Λ , then it can be expected that $\bar{R}(\delta_n^*, G_0)$ will converge to $R(G_0)$ and $\{\delta_n^*\}$ will be empirical Bayes.

The prior on \mathcal{S} is constructed following Rolph. Any such prior will do, although subjective considerations may motivate a particular choice of the h_i 's introduced below.

Let \mathcal{S} be the set of distribution functions on $[0, 1)$ and D be the subset of the infinite dimensional unit cube $[0, 1]^\infty$ whose elements are the possible

moment sequences for members of \mathcal{G} . Let D^N be the projection of D onto its first N coordinates. If $(m_1, \dots, m_N) \in D^N$ then there exist numbers $\underline{m}_{N+1} \leq \bar{m}_{N+1}$ (depending only on (m_1, \dots, m_N)) such that m_{N+1} is a possible $N + 1$ st moment if and only if $\underline{m}_{N+1} \leq m_{N+1} < \bar{m}_{N+1}$ when $\underline{m}_{N+1} < \bar{m}_{N+1}$ or $m_{N+1} = m_{N+1} = \bar{m}_{N+1}$ when $\underline{m}_{N+1} = \bar{m}_{N+1}$. Rolph constructs an "a priori" distribution on \mathcal{G} by working with the moments. A distribution μ^* is defined by its conditional densities on the i th moments and extended to the entire space of moments by the Kolmogorov extension theorem. If h_1, h_2, \dots are everywhere positive densities with respect to Lebesgue measure on $[0, 1)$ then μ^* is defined by

$$(4) \quad \mu^*(m | m_1, \dots, m_{i-1}) = h_i(m) / \int_{\underline{m}_i}^{\bar{m}_i} h_i(\bar{m}) d\bar{m} \quad \text{if } \underline{m}_i \leq m \leq \bar{m}_i \\ = 0 \quad \text{elsewhere.}$$

Since the moments of a distribution on $[0, 1)$ determine it uniquely, μ^* induces an a priori distribution μ on \mathcal{G} .

For a fixed $G_0 \in \mathcal{G}$ let X_1, X_2, \dots and X be a random sample with probability function f_{G_0} where $f_{G_0}(z) = \int_0^1 (1 - \lambda)\lambda^{z-1} dG_0(\lambda)$ for $z = 1, 2, \dots$. Let ν_0 be the measure induced by f_{G_0} on \mathbf{X} , the space of X and ν_0^∞ be the product measure induced by ν_0 on $(\mathbf{X} \times \mathbf{X} \times \dots)$. Let $\mu_{n,x}$ (which depends on $x^{(n)}$) denote the posterior distribution on \mathcal{G} given $X^{(n)} = x^{(n)}$ and $X = x$. Let μ_{G_0} be a point mass at G_0 . Then it follows from Theorem 6 of Rolph [6] that for each fixed x (except possibly a ν_0 null set) $\mu_{n,x} \rightarrow \mu_{G_0}$ a.o. ν_0^∞ in the sense that for every continuous function Φ on \mathcal{G}

$$(5) \quad \int_{\mathcal{G}} \Phi d\mu_{n,x} \rightarrow \int_{\mathcal{G}} \Phi dG_0,$$

where the topology on \mathcal{G} is the one induced by moment convergence.

Let X be a geometric random variable with parameter λ as in Example 1. Assume that λ is a random variable with distribution function G . For estimating λ with squared error loss the Bayes estimator δ_G is given by

$$(6) \quad \delta_G(x) = E_G(\Lambda | X = x) = f_G(x + 1) / f_G(x).$$

Now suppose that we are in the empirical Bayes situation with $\delta = \{\delta_n\}$ an empirical decision procedure. For a given δ_n it is easily seen that

$$(7) \quad \bar{R}(\delta_n, G) = \sum_{x^{(n)}} \int_0^1 \sum_x (\delta_n(x^{(n)}, x) - \lambda)^2 f_\lambda(x) dG(\lambda) f_G^n(x^{(n)}) \\ = \sum_{x^{(n)}} \sum_x (\delta_n(x^{(n)}, x) - f_G(x + 1) / f_G(x))^2 f_G(x) f_G^n(x^{(n)}) \\ + \sum_{x^{(n)}} \int_0^1 \sum_x (f_G(x + 1) / f_G(x) - \lambda)^2 f_\lambda(x) dG(\lambda) f_G^n(x^{(n)}).$$

Hence to choose $\delta^* = \{\delta_n^*\}$ which for each n minimizes the average, over \mathcal{G} , of $\bar{R}(\delta_n, G)$ with respect to μ it is enough to minimize the first term on the right-hand side of (7) term by term. For fixed $x^{(n)}$ and x this is done by defining $\delta^*(x^{(n)}, x)$ to be the expectation of $f_G(x + 1) / f_G(x)$ with respect to the posterior

distribution on \mathcal{G} , i.e.,

$$(8) \quad \delta_n^*(x^{(n)}, x) = \int_{\mathcal{G}} (f_G(x+1)/f_G(x)) d\mu_{n,x}.$$

Since μ puts positive probability on any open set of \mathcal{G} , δ_n^* is admissible. By (8), (5) and (6) we have that

$$(9) \quad \lim_{n \rightarrow \infty} \delta_n^*(x^{(n)}, x) = \delta_{G_0}(x) \quad \text{a.e.} \quad \nu_0^\infty.$$

Since the loss function is continuous and bounded we have by (9) and the Lebesgue dominated convergence theorem that $\lim_{n \rightarrow \infty} \bar{R}(\delta_n^*, G_0) = R(G_0)$. (For details see Section 3 of Robbins [5].) Hence $\delta^* = \{\delta_n^*\}$ is an admissible a.o. procedure.

In much the same way an admissible a.o. sequence can be constructed for the testing problem of Example 1 by noting that

$$(10) \quad \bar{R}(\delta_n, G) = \int_{\lambda^*}^1 (\lambda^* - \lambda) dG(\lambda) \\ + \sum_{x^{(n)}} \sum_x \delta_n(x^{(n)}, x) (\lambda^* f_G(x) - f_G(x+1)) f_G^n(x^{(n)}).$$

If $\delta^* = \{\delta_n^*\}$ is given by

$$\delta_n^*(x^{(n)}, x) = 1 \quad \text{if} \quad \int_{\mathcal{G}} (\lambda^* f_G(x) - f_G(x+1)) d\mu_{n,x} < 0 \\ = 0 \quad \text{otherwise,}$$

then δ_n^* minimizes the μ -average of (10) over \mathcal{G} . As before $\delta^* = \{\delta_n^*\}$ is an admissible a.o. sequence since for a given G the Bayes test δ_G is given by

$$\delta_G(x) = 1 \quad \text{if} \quad f_G(x+1)/f_G(x) > \lambda^* \\ = 0 \quad \text{if} \quad f_G(x+1)/f_G(x) < \lambda^*.$$

A more convenient expression for the estimate in (8) can be found as follows. Given $X^{(n)} = x^{(n)}$ and $X = x$ let n_j be the number of times j was observed. The sample can be written $(n_1, n_2, \dots, n_c, 0, \dots)$ where c is the largest observation. Since $f_G(x) = m_{x-1}(G) - m_x(G)$ where $m_j(G)$ denotes the j th moment of G the joint frequency function of (n_1, n_2, \dots) given n and m_1, \dots, m_N for $N \geq c$ is

$$f(n_1, \dots, n_c, \dots | n, m_1, \dots, m_N) = \binom{n+1}{n_1 \dots n_c} \prod_{x=1}^c (m_{x-1} - m_x)^{n_x}.$$

If μ^{N*} denotes the marginal density of μ^* on D^N then the posterior density of (m_1, \dots, m_N) given (n_1, \dots, n_c, \dots) on D^N is

$$g^N(m_1, \dots, m_N | n_1, \dots, n_c, \dots) \\ = \prod_{x=1}^c (m_{x-1} - m_x)^{n_x} \mu^{*N}(m_1, \dots, m_N) dm_1, \dots, dm_N / I(n_1, \dots, n_c, \dots)$$

where

$$I(n_1, \dots, n_c, \dots) \\ = \int_{D^N} \prod_{x=1}^c (m_{x-1} - m_x)^{n_x} \mu^{*N}(m_1, \dots, m_N) dm_1, \dots, dm_N.$$

It is easily seen that the estimate given in (9) is

$$(11) \quad \delta_n^*(x^{(n)}, x) = I(n_1', \dots, n_c', \dots) / I(n_1, \dots, n_c, \dots),$$

where $n_x' = n_x - 1$, $n_{x+1}' = n_{x+1} + 1$ and $n_j' = n_j$ for $j \neq x$ and $x + 1$.

The estimate in (11) is not difficult to compute if n and c are not too large and the h_i 's in (5) are simple. For example, if h_i is the uniform density on $[0, 1)$ for $i = 1, 2, \dots$ and $x^{(2)} = (1, 2)$ and $x = 1$ then

$$\delta_2^*((1, 2), 1) = I(1, 2, 0, \dots) / I(2, 1, 0, \dots) = \frac{1}{4}.$$

Acknowledgment. I wish to thank W. J. Hall, the associate editor, for some very helpful comments.

REFERENCES

- [1] JOHNS, M. V., JR. (1956). Contributions to the theory of empirical Bayes procedures in statistics. Doctoral thesis at Columbia Univ.
- [2] KARLIN, S. and RUBIN, H. (1956). The theory of decision procedures with monotone likelihood ratio. *Ann. Math. Statist.* **27** 272-299.
- [3] LEHMANN, E. L. (1959). *Testing Statistical Hypotheses*. Wiley, New York.
- [4] ROBBINS, H. (1955). An empirical Bayes approach to statistics. *Proc. Third Berkeley Symp. Math. Statist. Prob.* **1** 157-164.
- [5] ROBBINS, H. (1964). The empirical Bayes approach to statistical decision problems. *Ann. Math. Statist.* **35** 1-20.
- [6] ROLPH, J. (1968). Bayesian estimation of mixing distributions. *Ann. Math. Statist.* **39** 1289-1302.
- [7] SAMUEL, E. (1963). An empirical Bayes approach to the testing of certain parametric hypotheses. *Ann. Math. Statist.* **34** 1370-1385.