THE EMPIRICAL BAYES APPROACH TO STATISTICAL DECISION PROBLEMS[1]

By Herbert Robbins

Columbia University

1. Introduction. The empirical Bayes approach to statistical decision problems is applicable when the same decision problem presents itself repeatedly and independently with a fixed but unknown *a priori* distribution of the parameter. Not all decision problems in practice come to us imbedded in such a sequence, but when they do the empirical Bayes approach offers certain advantages over any approach which ignores the fact that the parameter is itself a random variable, as well as over any approach which assumes a personal or a conventional distribution of the parameter not subject to change with experience. My own interest in the empirical Bayes approach was renewed by recent work of E. Samuel [10], [11] and J. Neyman [6], to both of whom I am very much indebted. In keeping with the purpose of the Rietz Lecture I shall not confine myself to presenting new results and shall try to make the argument explicit at the risk of being tedious. In the current controversy between the Bayesian school and their opponents it is obvious that any theory of statistical inference will find itself in and out of fashion as the winds of doctrine blow. Here, then, are some remarks and references for further reading which I hope will interest my audience in thinking the matter through for themselves. Considerations of space have confined mention of the non-parametric case, and of the closely related "compound" approach in which no *a priori* distribution of the parameter is assumed, to the references at the end of the article.

2. The empirical Bayes decision problem. We begin by stating the kind of statistical decision problem with which we shall be concerned. This comprises

(a) A *parameter space* Λ with generic element λ. λ is the "state of nature" which is unknown to us.

(b) An *action space* A with generic element a.

(c) A *loss function* $L(a, \lambda) \geqq 0$ representing the loss we incur in taking action a when the parameter is λ.

(d) An *a priori distribution* G of λ on Λ. G may or may not be known to us.

(e) An *observable random variable* x belonging to a space X on which a σ-finite measure μ is defined. When the parameter is λ, x has a specified probability density f_λ with respect to μ.

The problem is to choose a *decision function* t, defined on X and with values in

Received 22 October 1963.

[1] Presented as the Rietz Lecture at the 26th Annual Meeting of the Institute of Mathematical Statistics, August 29, 1963, at Ottawa, Canada. Research sponsored by National Science Foundation Grant NSF-GP-316 at Columbia University.

A, such that when we observe x we shall take the action $t(x)$ and thereby incur the loss $L(t(x), \lambda)$. For any t the expected loss when λ is the parameter is

$$R(t, \lambda) = \int_X L(t(x), \lambda) f_\lambda(x) \, d\mu(x), \tag{1}$$

and hence the overall expected loss when the *a priori* distribution of λ is G is

$$R(t, G) = \int_\Lambda R(t, \lambda) \, dG(\lambda), \tag{2}$$

called the *Bayes risk* of t relative to G. We can write

$$R(t, G) = \int_X \phi_G(t(x), x) \, d\mu(x), \tag{3}$$

where we have set

$$\phi_G(a, x) = \int_\Lambda L(a, \lambda) f_\lambda(x) \, dG(\lambda). \tag{4}$$

To avoid needless complication we shall assume that there exists a decision function (d.f.) t_G such that for a.e. (μ) x,

$$\phi_G(t_G(x), x) = \min_a \phi_G(a, x). \tag{5}$$

Then for any d.f. t

$$R(t_G, G) = \int_X \min_a \phi_G(a, x) \, d\mu(x) \leqq R(t, G), \tag{6}$$

so that, defining

$$R(G) = R(t_G, G) = \int_X \phi_G(t_G(x), x) \, d\mu(x), \tag{7}$$

we have

$$R(G) = \min_t R(t, G). \tag{8}$$

Any d.f. t_G satisfying (5) minimizes the Bayes risk relative to G, and is called a *Bayes d.f.* relative to G. The functional R defined by (7) is called the *Bayes envelope functional* of G. When G is known we can use t_G and thereby incur the minimum possible Bayes risk $R(G)$.

There remains the problem of what to do when G is not known. To extreme Bayesians and extreme non-Bayesians this question will be empty, since to the former G will always be known, by introspection or otherwise, and to the latter G will not even exist. We shall, however, take the position that G exists, so that $R(t, G)$ is an appropriate criterion for the performance of any d.f. t, but that G is not known, and therefore t_G is not directly available to us.

Suppose now that the decision problem just described occurs repeatedly and independently, with the same unknown G throughout (for examples see [6]). Thus let

$$(\lambda_1, x_1), \qquad (\lambda_2, x_2), \qquad \cdots \tag{9}$$

be a sequence of pairs of random variables, each pair being independent of all the other pairs, the λ_n having a common *a priori* distribution G on Λ, and the conditional distribution of x_n given that $\lambda_n = \lambda$ being specified by the probability density f_λ. At the time when the decision about λ_{n+1} is to be made we have observed $x_1, \cdots, x_{n+1}$ (the values $\lambda_1, \lambda_2, \cdots$ remaining always unknown). We can therefore use for the decision about λ_{n+1} a function of x_{n+1} *whose form depends upon* $x_1, \cdots, x_n$; i.e. a function

$$t_n(\cdot) = t_n(x_1, \cdots, x_n; \cdot), \tag{10}$$

so that we shall take the action $t_n(x_{n+1}) \varepsilon A$ and thereby incur the loss $L(t_n(x_{n+1}), \lambda_{n+1})$. Our reason for doing this instead of using a fixed d.f. $t(\cdot)$ for each n (as might seem reasonable, since the successive problems are independent and have the same structure) is that we hope for large n to be able to extract some information about G from the values $x_1, \cdots, x_n$ which have been observed, hopefully in such a way that $t_n(\cdot)$ will be close to the optimal but unknown $t_G(\cdot)$ which we would use throughout if we knew G.

We therefore define an "empirical" or "adaptive" *decision procedure* to be a sequence $T = \{t_n\}$ of functions of the form (10) with values in A. For a given T, the expected loss on the decision about λ_{n+1}, *given* $x_1, \cdots, x_n$, will be (cf. (3))

$$\int_X \phi_G(t_n(x), x)\, d\mu(x), \tag{11}$$

and hence the overall expected loss will be

$$R_n(T, G) = \int_X E\phi_G(t_n(x), x)\, d\mu(x), \tag{12}$$

where E denotes expectation with respect to the n independent random variables $x_1, \cdots, x_n$ which have the common density with respect to μ on X given by

$$f_G(x) = \int_\Lambda f_\lambda(x)\, dG(\lambda), \tag{13}$$

the symbol x in (12) playing the role of a dummy variable of integration and not a random variable. From (5) and (12) it follows that always

$$R_n(T, G) \geqq R(G). \tag{14}$$

DEFINITION. If

$$\lim_{n\to\infty} R_n(T, G) = R(G) \tag{15}$$

we say that T is *asymptotically optimal* (*a.o.*) relative to G.

We now ask whether we can find a T which is in some sense "as good as possible" for large n relative to some class $\mathcal{G}$ of *a priori* distributions which we are willing to assume contains the true G. In particular, *can we find a T which is a.o. relative to every G in $\mathcal{G}$?* ($\mathcal{G}$ may be the class of all possible distributions on Λ.)

3. Some generalities on asymptotic optimality. Comparing (2.7) and (2.12) we see by Lebesgue's theorem on dominated convergence that for $T = \{t_n\}$ to be a.o. relative to a G it suffices that

$$\text{(A)} \qquad \lim_{n\to\infty} E\phi_G(t_n(x), x) = \phi_G(t_G(x), x) \qquad \text{(a.e. } (\mu)\ x),$$

and

$$\text{(B)} \quad E\phi_G(t_n(x), x) \leqq H(x)(\text{all } n), \qquad \text{where } \int_X H(x)\, d\mu(x) < \infty.$$

The main problem is (A); we shall summarily dispose of (B) by assuming that

$$\text{(C)} \qquad \int_\Lambda L(\lambda)\, dG(\lambda) < \infty,$$

where we have set

$$(1) \qquad 0 \leqq L(\lambda) = \sup_a L(a, \lambda) \leqq \infty.$$

Then setting

$$(2) \qquad H(x) = \int_\Lambda L(\lambda) f_\lambda(x)\, dG(\lambda) \geqq 0,$$

we have by (2.4) for any T,

$$(3) \qquad \phi_G(t_n(x), x) \leqq H(x) \qquad (\text{all} \quad n),$$

and by (C),

$$(4) \quad \int_X H(x)\, d\mu(x) = \int_\Lambda L(\lambda) \int_X f_\lambda(x)\, d\mu(x)\, dG(\lambda) = \int_\Lambda L(\lambda)\, dG(\lambda) < \infty,$$

and (3) and (4) imply that (B) holds. Moreover, from (4) it follows that

$$(5) \qquad H(x) < \infty \qquad (\text{a.e.} \quad (\mu)\ x)$$

and hence to prove that (A) holds it will suffice to prove that

$$\text{(D)} \qquad \text{p} \lim_{n\to\infty} \phi_G(t_n(x), x) = \phi_G(t_G(x), x) \qquad (\text{a.e.} \quad (\mu)\ x),$$

where by p lim we mean limit in probability. Hence (C) and (D) suffice to ensure that T is a.o. relative to G.

Let a_0 be an arbitrary fixed element of A and define

$$(6) \qquad \Delta_G(a, x) = \int_\Lambda [L(a, \lambda) - L(a_0, \lambda)] f_\lambda(x)\, dG(\lambda)$$

and

$$(7) \qquad L_0(x) = \int_\Lambda L(a_0, \lambda) f_\lambda(x)\, dG(\lambda),$$

so that under (C) we have, for a.e. (μ) x,

$$\phi_G(a, x) = L_0(x) + \Delta_G(a, x). \tag{8}$$

Suppose we can find a sequence of functions

$$\Delta_n(a, x) = \Delta_n(x_1, \cdots, x_n; a, x) \tag{9}$$

such that for a.e. (μ) x,

$$\text{p}\lim_{n\to\infty} \sup_a |\Delta_n(a, x) - \Delta_G(a, x)| = 0. \tag{10}$$

Let ϵ_n be any sequence of constants tending to 0 and set (subject to measurability conditions)

$$t_n(x) = t_n(x_1, \cdots, x_n; x) = \text{ any element } \bar{a} \varepsilon A \tag{11}$$

such that $\Delta_n(\bar{a}, x) \leqq \inf_a \Delta_n(a, x) + \epsilon_n$. Then by (2.5) and (8),

$$\begin{aligned} 0 &\leqq \Delta_G(t_n(x), x) - \Delta_G(t_G(x), x) \\ &= [\Delta_G(t_n(x), x) - \Delta_n(t_n(x), x)] + [\Delta_n(t_n(x), x) - \Delta_n(t_G(x), x)] \\ &\qquad + [\Delta_n(t_G(x), x) - \Delta_G(t_G(x), x)]. \end{aligned} \tag{12}$$

Given any $\epsilon > 0$ we have by (10) that for large n with probability as near 1 as we please the right hand side of (12) will be $\leqq \epsilon + \epsilon_n + \epsilon$; thus

$$\text{p}\lim_{n\to\infty} \Delta_G(t_n(x), x) = \Delta_G(t_G(x), x) \qquad (\text{a.e.} \quad (\mu)\ x), \tag{13}$$

which by (8) implies (D). We have therefore proved

THEOREM 1. *Let G be such that* (C) *holds, let $\Delta_n(a, x)$ be a sequence of functions of the form* (9) *and such that* (10) *holds, and define $T = \{t_n\}$ by* (11). *Then T is a.o. relative to G.*

When A is *finite* this yields

COROLLARY 1. *Let $A = \{a_0, \cdots, a_m\}$ be a finite set, let G be such that*

$$\int_\Lambda L(a_i, \lambda)\, dG(\lambda) < \infty \qquad (i = 0, \cdots, m), \tag{14}$$

and let $\Delta_{i,n}(x) = \Delta_{i,n}(x_1, \cdots, x_n; x)$ for $i = 1, \cdots, m$ and $n = 1, 2, \cdots$ be such that for a.e. (μ) x,

$$\text{p}\lim_{n\to\infty}\Delta_{i,n}(x) = \int_\Lambda [L(a_i, \lambda) - L(a_0, \lambda)] f_\lambda(x)\, dG(\lambda). \tag{15}$$

Set $\Delta_{0,n}(x) = 0$ and define

$$\begin{aligned} t_n(x) &= a_k, \qquad \textit{where } k \textit{ is any integer } 0 \leqq k \leqq m \textit{ such that} \\ &\Delta_{k,n}(x) = \min [0, \Delta_{1,n}(x), \cdots, \Delta_{m,n}(x)]. \end{aligned} \tag{16}$$

Then $T = \{t_n\}$ is a.o. relative to G.

In the important case $m = 1$ (hypothesis testing) this becomes

COROLLARY 2. *Let* $A = \{a_0, a_1\}$, *let* G *be such that*

$$\int_\Lambda L(a_i, \lambda)\, dG(\lambda) < \infty \qquad (i = 0, 1), \tag{17}$$

and let $\Delta_n(x) = \Delta_n(x_1, \cdots, x_n; x)$ *be such that for a.e.* (μ) x,

$$\text{p}\lim_{n\to\infty} \Delta_n(x) = \Delta_G(x) = \int_\Lambda [L(a_1, \lambda) - L(a_0, \lambda)] f_\lambda(x)\, dG(\lambda). \tag{18}$$

Define

$$\begin{aligned} t_n(x) &= a_0, & &\textit{if} \quad \Delta_n(x) \geqq 0, \\ &= a_1, & &\textit{if} \quad \Delta_n(x) < 0. \end{aligned} \tag{19}$$

Then $T = \{t_n\}$ *is a.o. relative to* G.

We proceed to give an example in which a sequence $\Delta_n(x)$ satisfying (18) can be constructed.

4. The Poisson case. We consider the problem of testing a one-sided null hypothesis $H_0 : \lambda \leqq \lambda^*$ concerning the value of a Poisson parameter λ. Thus let $\Lambda = \{0 < \lambda < \infty\}$, $A = \{a_0, a_1\}$, where a_0 = "accept H_0", a_1 = "reject H_0" and

$$\begin{aligned} X &= \{0, 1, 2, \cdots\}, \qquad \mu = \text{counting measure on } X, \\ f_\lambda(x) &= e^{-\lambda}\lambda^x/x!. \end{aligned} \tag{1}$$

It remains to specify the loss functions $L(a_i, \lambda)$; we shall take them to be

$$\begin{aligned} L(a_0, \lambda) &= 0 & &\text{if} \quad \lambda \leqq \lambda^*, \\ &= \lambda - \lambda^* & &\text{if} \quad \lambda \geqq \lambda^*, \\ L(a_1, \lambda) &= \lambda^* - \lambda & &\text{if} \quad \lambda \leqq \lambda^*, \\ &= 0 & &\text{if} \quad \lambda \geqq \lambda^*. \end{aligned} \tag{2}$$

Thus (very conveniently)

$$L(a_1, \lambda) - L(a_0, \lambda) = \lambda^* - \lambda \qquad (0 < \lambda < \infty), \tag{3}$$

and

$$\Delta_G(x) = \int_\Lambda [L(a_1, \lambda) - L(a_0, \lambda)] f_\lambda(x)\, dG(\lambda) = \int_0^\infty (\lambda^* - \lambda) \frac{e^{-\lambda}\lambda^x}{x!}\, dG(\lambda). \tag{4}$$

Now by (2.13),

$$f_G(x) = P(x_j = x) = \int_0^\infty \frac{e^{-\lambda}\lambda^x}{x!}\, dG(\lambda), \tag{5}$$

so that we can write

$$\Delta_G(x) = \lambda^* f_G(x) - (x + 1) f_G(x + 1). \tag{6}$$

Define

$$\delta(x, y) = 1 \quad \text{if } x = y, \qquad = 0 \quad \text{if } x \neq y, \tag{7}$$

and consider the expression

$$u_n(x) = u_n(x_1, \cdots, x_n; x) = n^{-1} \sum_{j=1}^{n} \delta(x, x_j). \tag{8}$$

Noting that

$$E\delta(x, x_j) = P(x_j = x) = f_G(x), \tag{9}$$

it follows from the law of large numbers that

$$\text{p} \lim_{n\to\infty} u_n(x) = f_G(x) \qquad (x = 0, 1, \cdots), \tag{10}$$

and hence that, setting

$$\Delta_n(x) = \lambda^* u_n(x) - (x + 1)u_n(x + 1), \tag{11}$$

we have for $x = 0, 1, \cdots$

$$\text{p} \lim_{n\to\infty} \Delta_n(x) = \lambda^* f_G(x) - (x + 1)f_G(x + 1) = \Delta_G(x). \tag{12}$$

Setting

$$t_n(x) = a_0 \quad \text{if } \lambda^* u_n(x) - (x + 1)u_n(x + 1) \geqq 0, \qquad = a_1 \quad \text{otherwise}, \tag{13}$$

it follows from Corollary 2 of Section 3 that T *is a.o. relative to every* G *such that*

$$\int_0^\infty \lambda \, dG(\lambda) < \infty. \tag{14}$$

We remark that we could equally well have defined $u_n(x)$ to be

$$u_n(x) = [1/(n + 1)] \sum_{j=1}^{n+1} \delta(x, x_j), \tag{15}$$

since (10) holds as well for (15) as for (8). Using (15) the corresponding T would require us to take action a_0 on λ_n if and only if

$$\lambda^* \geqq \frac{(x_n + 1)(\text{number of terms } x_1, \cdots, x_n \text{ equal to } x_n + 1)}{(\text{number of terms } x_1, \cdots, x_n \text{ equal to } x_n)}. \tag{16}$$

For the problem of point estimation with squared error loss it was shown by M. V. Johns, Jr. [2] that the right hand side of (16) is in fact an a.o. point estimator of λ_n for every G such that

$$\int_0^\infty \lambda^2 \, dG(\lambda) < \infty. \tag{17}$$

The much easier result of the present section on hypothesis testing uses a loss structure (2) suggested by Johns in [4].

The relation (6) was basic to our construction (11) of a sequence $\Delta_n(x)$ satisfying (3.18); now, (6) is a special property of the Poisson distribution (1) and the loss structure (2), and therefore it might seem that the application of Corollary 2 to empirical Bayes hypothesis testing would be very limited. Such is not the case. The application of Corollary 2 to more general loss structures, and to many of the most common discrete and continuous parametric distributions of statistics is discussed in [4], [9], [11]. Instead of giving a review of these results here, we shall consider a case in which no asymptotically optimal T exists but the empirical Bayes approach is still useful.

5. An example in which asymptotic optimality does not exist but all is not lost. Let x be a random variable with only two values, 0 and 1, with respective probabilities $1 - \lambda$ and λ, the unknown parameter λ lying in the interval $\Lambda = \{0 \leqq \lambda \leqq 1\}$. On the basis of a single observation of x we want to estimate λ; if our estimate is $a \varepsilon A = \{0 \leqq a \leqq 1\}$ the loss will be taken to be $L(a, \lambda) = (\lambda - a)^2$. A d.f. t is determined by the two constants $t(0)$, $t(1)$ which are at our disposal on the unit interval A; the expected loss in using t for a given λ is

$$(1)\qquad \begin{aligned} R(t, \lambda) &= (1 - \lambda)(\lambda - t(0))^2 + \lambda(\lambda - t(1))^2 \\ &= t^2(0) + [t^2(1) - 2t(0) - t^2(0)]\lambda + [1 - 2t(1) + 2t(0)]\lambda^2. \end{aligned}$$

Consider the particular family of d.f.'s t_α defined for $0 < \alpha < 1$ by setting

$$(2)\qquad t_\alpha(0) = \tfrac{1}{2}\alpha, \qquad t_\alpha(1) = \tfrac{1}{2}(1 + \alpha).$$

It is easily seen from (1) that

$$(3)\qquad R(t_\alpha, \lambda) = \tfrac{1}{4}[\alpha^2 + (1 - 2\alpha)\lambda].$$

For $\alpha = \frac{1}{2}$ we shall denote t_α by t^*, so that

$$(4)\qquad t^*(0) = \tfrac{1}{4}, \qquad t^*(1) = \tfrac{3}{4}, \qquad R(t^*, \lambda) \equiv \tfrac{1}{16}.$$

For any *a priori* distribution G of λ let

$$(5)\qquad \nu_i = \nu_i(G) = \int_0^1 \lambda^i \, dG(\lambda) \qquad (i = 1, 2).$$

Then from (1) it follows that for any d.f. t,

$$(6)\qquad \begin{aligned} R(t, G) &= \int_0^1 R(t, \lambda)\, dG(\lambda) \\ &= t^2(0) + \nu_1[t^2(1) - 2t(0) - t^2(0)] + \nu_2[1 - 2t(1) + 2t(0)]. \end{aligned}$$

Apart from the trivial cases in which $\nu_1 = 0$ or 1 we have after some simple algebra the formula

$$(7)\qquad \begin{aligned} R(t, G) &= (\nu_1 - \nu_2)(\nu_2 - \nu_1^2)/\nu_1(1 - \nu_1) \\ &\quad + (1 - \nu_1)[t(0) - (\nu_1 - \nu_2)/(1 - \nu_1)]^2 + \nu_1[t(1) - \nu_2/\nu_1]^2, \end{aligned}$$

from which it follows that for a given G, $R(t, G)$ is minimized uniquely by the Bayes d.f. t_G for which

$$t_G(0) = (\nu_1 - \nu_2)/(1 - \nu_1), \qquad t_G(1) = \nu_2/\nu_1, \tag{8}$$

with

$$R(G) = R(t_G, G) = (\nu_1 - \nu_2)(\nu_2 - \nu_1^2)/\nu_1(1 - \nu_1). \tag{9}$$

Each t_α (and in particular t^*) is a Bayes d.f.; it suffices to find a distribution of λ such that

$$(\nu_1 - \nu_2)/(1 - \nu_1) = \tfrac{1}{2}\alpha, \qquad \nu_2/\nu_1 = (1 + \alpha)/\alpha, \tag{10}$$

and this is provided e.g. by the distribution G_α with the density

$$[B(\alpha, 1 - \alpha)]^{-1}\lambda^{\alpha-1}(1 - \lambda)^{(1-\alpha)-1} \tag{11}$$

for which

$$\nu_1 = \alpha, \qquad \nu_2 = \tfrac{1}{2}\alpha(1 + \alpha), \qquad R(G_\alpha) = \tfrac{1}{4}\alpha(1 - \alpha). \tag{12}$$

The fact that t^* is the Bayes d.f. relative to $G_{\frac{1}{2}}$ and that for any G,

$$R(t^*, G) = \int_0^1 R(t^*, \lambda)\, dG(\lambda) = \tfrac{1}{16} \tag{13}$$

has the important consequence that

$$\sup_G R(t, G) > \tfrac{1}{16} \qquad \text{for every} \quad t \neq t^*. \tag{14}$$

For if for some t', $\sup_G R(t', G) \leqq \frac{1}{16}$, then in particular

$$\tfrac{1}{16} = R(t^*, G_{\frac{1}{2}}) \leqq R(t', G_{\frac{1}{2}}) \leqq \tfrac{1}{16}, \tag{15}$$

so that

$$R(t', G_{\frac{1}{2}}) = R(t^*, G_{\frac{1}{2}}) = \tfrac{1}{16} \tag{16}$$

and therefore $t' = t^*$. Thus t^* is the unique "minimax" d.f. in the sense that *it minimizes the maximum Bayes risk relative to the class of all a priori distributions* G. When nothing is known about G it is therefore not unreasonable (the avoidance of a direct endorsement of any decision function is customary in the literature) to use t^*; the Bayes risk will then be $\frac{1}{16}$ irrespective of G, while for any $t \neq t^*$ the Bayes risk will be $> \frac{1}{16}$ for some G (in particular for any G with $\nu_1 = \frac{1}{2}$, $\nu_2 = \frac{3}{8}$; e.g. $G_{\frac{1}{2}}$).

For any $0 < \alpha < 1$ let $\mathcal{G}_\alpha$ denote the class of all G such that $\nu_1(G) = \alpha$. For any G in $\mathcal{G}_\alpha$ (in particular for G_α) we see from (2) and (7) after a little algebra that

$$R(t_\alpha, G) = \tfrac{1}{4}\alpha(1 - \alpha) \qquad (G \varepsilon \mathcal{G}_\alpha) \tag{17}$$

irrespective of the value of $\nu_2(G)$. It therefore follows as above that

$$\sup_{G \varepsilon \mathcal{G}_\alpha} R(t, G) > \tfrac{1}{4}\alpha(1 - \alpha) \qquad \text{for every} \quad t \neq t_\alpha, \tag{18}$$

so that relative to the class $\mathcal{G}_\alpha$, t_α is the unique minimax d.f. in the sense that *it minimizes the maximum Bayes risk relative to the class* $\mathcal{G}_\alpha$. If nothing is known about G except that $\nu_1(G) = \alpha$ it is therefore not unreasonable to use t_α; the Bayes risk will then be $\frac{1}{4}\alpha(1 - \alpha)$, while for any other d.f. the Bayes risk will be $> \frac{1}{4}\alpha(1 - \alpha)$ for some G in $\mathcal{G}_\alpha$ (in particular for any G with $\nu_1 = \alpha$, $\nu_2 = \frac{1}{2}\alpha(1 + \alpha)$; e.g. G_α).

It follows from the above (or can be verified directly) that

$$(\nu_1 - \nu_2)(\nu_2 - \nu_1^2)/\nu_1(1 - \nu_1) \leqq \tfrac{1}{4}\nu_1(1 - \nu_1) \leqq \tfrac{1}{16}; \tag{19}$$

equality holding respectively when $\nu_2 = \frac{1}{2}\nu_1(1 + \nu_1)$ and when $\nu_1 = \frac{1}{2}$.

Suppose now that we confront this decision problem repeatedly with an unknown G. The sequence $x_1, x_2, \cdots$ is an independent and identically distributed sequence of 0's and 1's with

$$P(x_i = 1) = \int_0^1 \lambda \, dG(\lambda) = \nu_1(G), \qquad P(x_i = 0) = 1 - \nu_1(G); \tag{20}$$

thus the distribution of the x_i depends only on $\nu_1(G)$. Since t_G, defined by (8), involves $\nu_2(G)$ as well, it follows that no T can be a.o. relative to every G in a class $\mathcal{G}$ unless ν_2 is a function of ν_1 in $\mathcal{G}$, which is not likely to be the case in practice.

On the other hand, let

$$u_n = (1/n) \sum_{i=1}^{n} x_i, \tag{21}$$

and consider the decision procedure $\tilde{T} = \{t_n\}$ with

$$t_n(0) = \tfrac{1}{2}u_n, \qquad t_n(1) = \tfrac{1}{2}(1 + u_n). \tag{22}$$

For any G in $\mathcal{G}_\alpha$, by the law of large numbers, $u_n \to \alpha$ and hence $t_n \to t_\alpha$ with probability 1 as $n \to \infty$. In fact, since

$$Ex_i = \alpha = Ex_i^2, \qquad \operatorname{Var} x_i = \alpha(1 - \alpha), \tag{23}$$

it follows that

$$Eu_n = \alpha, \qquad Eu_n^2 = \operatorname{Var} u_n + \alpha^2 = \alpha(1 - \alpha)/n + \alpha^2 \tag{24}$$

and hence from (6) that

$$\begin{aligned} &R_n(\tilde{T}, G) \\ &= E[\tfrac{1}{4}u_n^2 + \alpha\{\tfrac{1}{4}(1 + 2u_n + u_n^2) - u_n - \tfrac{1}{4}u_n^2\} + \nu_2(1 - 1 - u_n + u_n)] \\ &= \tfrac{1}{4}E[u_n^2 - 2u_n + \alpha] = \tfrac{1}{4}\alpha(1 - \alpha)[1 + 1/n] = R(t_\alpha, G)[1 + 1/n]. \end{aligned} \tag{25}$$

Thus for large n we will do almost as well by using $\tilde{T}$ *as we could do if we knew* $\nu_1(G) = \alpha$ *and used* t_α. We have in fact for $G \,\varepsilon\, \mathcal{G}_\alpha$,

$$R_n(\tilde{T}, G) - R(t_\alpha, G) = \alpha(1 - \alpha)/4n \leqq 1/16n, \tag{26}$$

while

$$(27)\quad \begin{aligned} R_n(\tilde{T}, G) - R(t^*, G) &= \alpha(1 - \alpha)(n + 1)/4n - \tfrac{1}{16} \\ &= -[\tfrac{1}{4}(1 - 2\alpha)]^2 + \alpha(1 - \alpha)/4n. \end{aligned}$$

This example illustrates the fact that even when an a.o. T does not exist, or when it does exist but $R_n(T, G)$ is too slowly convergent to $R(G)$, it may be worthwhile to use a T which is at least "asymptotically subminimax." R. Cogburn has work in progress in this direction for the case in which x has a general binomial distribution (for which see also [9] and [11]). The general problem of finding "good" T's has hardly been touched, and efforts along this line should yield interesting and useful results.

6. Estimating the *a priori* distribution: the general case. Returning to the general formulation of Sections 2 and 3 and confining ourselves for simplicity to the case $A = \{a_0, a_1\}$ and $\Lambda = \{-\infty < \lambda < \infty\}$, we recall that an a.o. T exists relative to the class $\mathcal{G}$ defined by (3.17) whenever we can find a sequence $\Delta_n(x) = \Delta_n(x_1, \cdots, x_n; x)$ such that for a.e. $(\mu)x$,

$$(1)\quad \text{p}\lim_{n\to\infty} \Delta_n(x) = \Delta_G(x) = \int_{-\infty}^{\infty} [L(a_1, \lambda) - L(a_0, \lambda)]f_\lambda(x)\, dG(\lambda)$$

for every G in $\mathcal{G}$. One way to construct such a sequence (not the one used in Section 4) is to find a sequence $G_n(\lambda) = G_n(x_1, \cdots x_n; \lambda)$ of random distribution functions in λ such that

$$(2)\quad P[\lim_{n\to\infty} G_n(\lambda) = G(\lambda) \text{ at every continuity point } \lambda \text{ of } G] = 1.$$

If we have such a sequence G_n of random estimators of G then we can set

$$(3)\quad \Delta_n(x) = \int_{-\infty}^{\infty} [L(a_1, \lambda) - L(a_0, \lambda)]f_\lambda(x)\, dG_n(\lambda);$$

if for a.e. (μ) fixed x the function

$$(4)\quad [L(a_1, \lambda) - L(a_0, \lambda)]f_\lambda(x)$$

is continuous and bounded in λ, then the Helly-Bray theorem guarantees (1).

We shall now describe one method—a special case of the "minimum distance" method of J. Wolfowitz—for constructing a particular sequence $G_n(\lambda)$ of random estimators of an unknown G, and then prove a theorem which ensures that under appropriate conditions on the family $f_\lambda(x)$ the relation (2) will hold for any G whatever.

In doing this we shall relax the condition that the distribution of x for given λ is given in terms of a density f_λ with respect to a measure μ, and instead assume only that for every $\lambda \varepsilon \Lambda = \{-\infty < \lambda < \infty\}$, $F_\lambda(x)$ is a specified distribution function in x, and for every fixed $x \varepsilon X = \{-\infty < x < \infty\}$, $F_\lambda(x)$ is a Borel measurable function of λ. (A function $F(x)$ defined for $-\infty < x < \infty$ will be called a distribution function if F is non-decreasing, continuous on the right, and $\lim_{x\to-\infty} F(x) = 0$, $\lim_{x\to\infty} F(x) = 1$.)

For any distribution function G of λ we define

$$F_G(x) = \int_{-\infty}^{\infty} F_\lambda(x)\, dG(\lambda); \tag{5}$$

then F_G is a distribution function in x.

Let $x_1, x_2, \cdots$ be a sequence of independent random variables with F_G as their common distribution function, and define

$$B_n(x) = B_n(x_1, \cdots, x_n; x) = (\text{no. of terms } x_1, \cdots, x_n \text{ which are } \leqq x)/n. \tag{6}$$

For any two distribution functions F_1, F_2 define the distance

$$\rho(F_1, F_2) = \sup_x |F_1(x) - F_2(x)|, \tag{7}$$

and let ϵ_n be any sequence of constants tending to 0.

Let $\mathcal{G}$ be any class of distribution functions in λ which contains G, and define

$$d_n = \inf_{\bar{G}\varepsilon\mathcal{G}} \rho(B_n, F_{\bar{G}}). \tag{8}$$

Let $G_n(\lambda) = G_n(x_1, \cdots, x_n; \lambda)$ be any element of $\mathcal{G}$ such that

$$\rho(B_n, F_{G_n}) \leqq d_n + \epsilon_n. \tag{9}$$

We say that the sequence G_n so defined is *effective* for $\mathcal{G}$ if (2) holds for every G in $\mathcal{G}$. We now state

THEOREM 2. *Assume that*

(A) *For every fixed x, $F_\lambda(x)$ is a continuous function of λ.*

(B) *The limits $F_{-\infty}(x) = \lim_{\lambda\to-\infty} F_\lambda(x)$, $F_\infty(x) = \lim_{\lambda\to\infty} F_\lambda(x)$ exist for every x.*

(C) *Neither $F_{-\infty}$ nor F_∞ is a distribution function.*

(D) *If G_1, G_2 are any two distribution functions in λ such that $F_{G_1} = F_{G_2}$, then $G_1 = G_2$.*

Then the sequence G_n defined by (9) is effective for the class $\mathcal{G}$ of all distribution functions in λ.

PROOF. By the Glivenko-Cantelli theorem,

$$P[\lim_{n\to\infty} \rho(B_n, F_G) = 0] = 1. \tag{10}$$

Since

$$\begin{aligned} \rho(F_{G_n}, F_G) &\leqq \rho(F_{G_n}, B_n) + \rho(B_n, F_G) \\ &\leqq d_n + \epsilon_n + \rho(B_n, F_G) \leqq \rho(B_n, F_G) + \epsilon_n + \rho(B_n, F_G), \end{aligned} \tag{11}$$

it follows from (10) that with probability 1 the sequence $x_1, x_2, \cdots$ is such that

$$\lim_{n\to\infty} \int_{-\infty}^{\infty} F_\lambda(x)\, dG_n(\lambda) = \int_{-\infty}^{\infty} F_\lambda(x)\, dG(\lambda) \qquad \text{(uniformly in } x). \tag{12}$$

Now consider any fixed sequence $x_1, x_2, \cdots$ such that (12) holds and let G_{k_n} be any subsequence of G_n such that $G_{k_n}(\lambda) \to G^*(\lambda)$ at every continuity point λ of G^*, where G^* is a distribution function in the weak sense, $0 \leqq G^*(-\infty)$, $G^*(\infty) \leqq 1$. By a simple extension of the Helly-Bray theorem it follows from (A) and (B) that for every x,

$$\lim_{n\to\infty} \int_{-\infty}^{\infty} F_\lambda(x)\, dG_{k_n}(\lambda) = \int_{-\infty}^{\infty} F_\lambda(x)\, dG^*(\lambda) + G^*(-\infty)F_{-\infty}(x) + [1 - G^*(\infty)]F_\infty(x), \tag{13}$$

and hence from (12) that for every x,

$$\int_{-\infty}^{\infty} F_\lambda(x)\, dG(\lambda) = \int_{-\infty}^{\infty} F_\lambda(x)\, dG^*(\lambda) + G^*(-\infty)F_{-\infty}(x) + [1 - G^*(\infty)]F_\infty(x). \tag{14}$$

If we can show that $G^*(-\infty) = 0$ and $G^*(\infty) = 1$ then it will follow from (D) that $G = G^*$, and hence, since G^* denoted the weak limit of *any* convergent subsequence of G_n, that (2) holds. We shall complete the proof of the theorem by showing that (C) implies that $G^*(-\infty) = 0$ and $G^*(\infty) = 1$.

Since $F_{-\infty}$ is the limit as $\lambda \to -\infty$ of F_λ, it is a nondecreasing function of x such that

$$0 \leqq F_{-\infty}(-\infty), \qquad F_{-\infty}(\infty) \leqq 1, \tag{15}$$

and similarly for F_∞. Let $x \to -\infty$ in (14). By Lebesgue's theorem of bounded convergence,

$$0 = G^*(-\infty)F_{-\infty}(-\infty) + [1 - G^*(\infty)]F_\infty(-\infty). \tag{16}$$

Hence if $G^*(-\infty) \neq 1$ then $F_{-\infty}(-\infty) = 0$, and if $G^*(\infty) \neq 1$ then $F_\infty(-\infty) = 0$. Similarly, by letting $x \to \infty$ in (14) we see that if $G^*(-\infty) \neq 0$ then $F_{-\infty}(\infty) = 1$, and if $G^*(\infty) \neq 1$ then $F_\infty(\infty) = 1$. Suppose now that a_n is any sequence of constants converging to a limit a from the right. Then from (14), putting $x = a_n$, letting $n \to \infty$, and subtracting (14) for $x = a$, we see that

$$G^*(-\infty)[F_{-\infty}(a + 0) - F_{-\infty}(a)] + [1 - G^*(\infty)][F_\infty(a + 0) - F_\infty(a)] = 0. \tag{17}$$

Hence if $G^*(-\infty) \neq 0$ then $F_{-\infty}(a + 0) = F_{-\infty}(a)$, and if $G^*(\infty) \neq 1$ then $F_\infty(a + 0) = F_\infty(a)$. It follows that if $G^*(-\infty) \neq 0$ then $F_{-\infty}$ is a distribution function and if $G^*(\infty) \neq 1$ then F_∞ is a distribution function. Hence by (C), $G^*(-\infty) = 0$ and $G^*(\infty) = 1$. This completes the proof.

EXAMPLE 1. (location parameter) Let F be a continuous distribution function (e.g., the normal distribution function) with a characteristic function which never vanishes,

$$\phi_F(t) = \int_{-\infty}^{\infty} e^{ixt}\, dF(x) \neq 0 \qquad \text{(all } t\text{)}. \tag{18}$$

Set $F_\lambda(x) = F(x - \lambda)$; then (A), (B), (C) hold. If G_1, G_2 are any two distribution functions such that $F_{G_1} = F_{G_2}$; i.e., such that

$$\int_{-\infty}^{\infty} F(x - \lambda)\, dG_1(\lambda) = \int_{-\infty}^{\infty} F(x - \lambda)\, dG_2(\lambda) \qquad \text{(all } x\text{)}, \tag{19}$$

then

$$\phi_F(t)\phi_{G_1}(t) = \phi_F(t)\phi_{G_2}(t) \qquad \text{(all } t\text{)}, \tag{20}$$

and hence

$$\phi_{G_1}(t) = \phi_{G_2}(t) \qquad \text{(all } t\text{)}, \tag{21}$$

so that $G_1 = G_2$. Hence (D) holds, and by Theorem 1 the sequence G_n defined by (9) is effective for the class $\mathcal{G}$ of all distributions G of λ.

When the parameter space Λ is not the whole line, the statement and proof of Theorem 2 can be appropriately modified. As an example, suppose $\Lambda = \{0 \leqq \lambda < \infty\}$. Then we can prove in exactly the same way as for Theorem 2

THEOREM 3. *Assume that*

(A) *As in Theorem* 2.

(B) *The limit* $F_\infty(x) = \lim_{\lambda\to\infty} F_\lambda(x)$ *exists for every* x.

(C) F_∞ *is not a distribution function.*

(D) *If* G_1, G_2 *are any two distribution functions in* λ *which assign unit probability to* $\Lambda = \{0 \leqq \lambda < \infty\}$ *such that*

$$\int_\Lambda F_\lambda(x)\, dG_1(\lambda) = \int_\Lambda F_\lambda(x)\, dG_2(\lambda) \qquad \text{(all } x\text{)},$$

then $G_1 = G_2$.

Then the sequence G_n *defined by* (9) *is effective for the class* $\mathcal{G}$ *of all distributions which assign unit probability to* Λ.

EXAMPLE 2. (Poisson parameter) Let

$$F_0(x) = 0 \quad \text{for } x < 0, \qquad = 1 \quad \text{for } x \geqq 0, \tag{22}$$

and for $0 < \lambda < \infty$ let

$$F_\lambda(x) = \sum_{0 \leqq i \leqq x} e^{-\lambda}\lambda^i/i!. \tag{23}$$

Then (A), (B), and (C) hold.

Let $G \varepsilon \mathcal{G}$; then

$$F_G(x) = \int_\Lambda F_\lambda(x)\, dG(\lambda) = \sum_{0 \leqq i \leqq x} \int_\Lambda f_\lambda(i)\, dG(\lambda), \tag{24}$$

where

$$\begin{aligned} f_0(i) &= 1 && \text{for } i = 0, \\ &= 0 && \text{for } i = 1, 2, \cdots \\ f_\lambda(i) &= e^{-\lambda}\lambda^i/i! && \text{for } i = 0, 1, \cdots \text{ and } 0 < \lambda < \infty. \end{aligned} \tag{25}$$

Now

$$\begin{aligned} F_G(0) &= \int_\Lambda f_\lambda(0)\, dG(\lambda) \\ F_G(n) - F_G(n-1) &= \int_\Lambda f_\lambda(n)\, dG(\lambda) \qquad (n = 1, 2, \cdots), \end{aligned} \tag{26}$$

so if $F_{G_1} = F_{G_2}$ then

$$\int f_\lambda(n)\, dG_1(\lambda) = \int f_\lambda(n)\, dG_2(\lambda) \qquad (n = 0, 1, \cdots). \tag{27}$$

Define the set functions

$$H_j(B) = \int_B e^{-\lambda}\, dG_j(\lambda) \Big/ \int_\Lambda e^{-\lambda}\, dG_j(\lambda) \qquad (j = 1, 2); \tag{28}$$

then H_j is a probability measure on the Borel sets. Since by (27),

$$c = \int_\Lambda e^{-\lambda}\, dG_1(\lambda) = \int_\Lambda f_\lambda(0)\, dG_1(\lambda) = \int_\Lambda f_\lambda(0)\, dG_2(\lambda) = \int_\Lambda e^{-\lambda}\, dG_2(\lambda), \tag{29}$$

we can write

$$H_j(B) = \frac{1}{c} \int_B e^{-\lambda}\, dG_j(\lambda) \qquad (j = 1, 2) \tag{30}$$

where $0 < c < \infty$. Since

$$dH_j/dG_j = e^{-\lambda}/c \tag{31}$$

we have for $n = 1, 2, \cdots$ and $j = 1, 2$

$$\int_\Lambda \lambda^n\, dH_j(\lambda) = \frac{1}{c} \int_\Lambda e^{-\lambda} \lambda^n\, dG_j(\lambda) = \frac{n!}{c} \int_\Lambda f_\lambda(n)\, dG_j(\lambda), \tag{32}$$

so that by (27)

$$\alpha_n = \int_\Lambda \lambda^n\, dH_1(\lambda) = \int_\Lambda \lambda^n\, dH_2(\lambda) \qquad (n = 1, 2, \cdots); \tag{33}$$

moreover, since $0 \leqq e^{-\lambda} \lambda^n \leqq n!$ for $0 \leqq \lambda < \infty$, we have

$$0 \leqq \alpha_n = \frac{1}{c} \int_\Lambda e^{-\lambda} \lambda^n\, dG_j(\lambda) \leqq \frac{n!}{c} \qquad (n = 1, 2, \cdots), \tag{34}$$

so that the series $\sum_1^\infty (\alpha_n/n!)(\frac{1}{2})^n < \infty$. From a theorem of H. Cramér (*Mathematical Methods of Statistics*, p. 176) it follows that $H_1 = H_2$. Since

$$G_j(B) = \int_B \frac{dG_j}{dH_j}\, dH_j(\lambda) = \int_B c e^{\lambda}\, dH_j(\lambda) \qquad (j = 1, 2) \tag{35}$$

it follows that $G_1 = G_2$, so that (D) holds.

We conclude with an example in which $\Lambda = \{0 < \lambda < \infty\}$, 0 here playing the role of $-\infty$ in Theorem 2.

EXAMPLE 3. (uniform distribution) Define for $\lambda \varepsilon \Lambda = \{0 < \lambda < \infty\}$

$$(36) \qquad \begin{aligned} F_\lambda(x) &= 0 && \text{for } x \leqq 0, \\ &= x/\lambda && \text{for } 0 < x < \lambda, \\ &= 1 && \text{for } x \geqq \lambda. \end{aligned}$$

Then

$$(37) \qquad \begin{aligned} \lim_{\lambda \to 0} F_\lambda(x) &= 0 && \text{for } x \leqq 0, \\ &= 1 && \text{for } x > 0 \end{aligned}$$

and

$$(38) \qquad \lim_{\lambda \to \infty} F_\lambda(x) \equiv 0$$

are not distribution functions, and (A), (B), (C) hold.

For any G which assigns unit probability to Λ we have for $x > 0$,

$$(39) \qquad F_G(x) = \int_\Lambda F_\lambda(x)\, dG(\lambda) = \int_{\{0<\lambda\leqq x\}} 1 \cdot dG(\lambda) + x \int_{\{\lambda > x\}} \frac{dG(\lambda)}{\lambda} .$$

Hence if $F_{G_1} = F_{G_2}$ then

$$(40) \qquad G_1(x) + x \int_{\{\lambda > x\}} \frac{dG_1(\lambda)}{\lambda} = G_2(x) + x \int_{\{\lambda > x\}} \frac{dG_2(\lambda)}{\lambda} .$$

If x is any common continuity point of G_1 and G_2 then

$$(41) \quad \int_{\{\lambda > x\}} \frac{dG_j(\lambda)}{\lambda} = \left[\frac{G_j(\lambda)}{\lambda}\right]_x^\infty + \int_{\{\lambda > x\}} \frac{G_j(\lambda)}{\lambda^2}\, d\lambda = -\frac{G_j(x)}{x} + \int_{\{\lambda > x\}} \frac{G_j(\lambda)}{\lambda^2}\, d\lambda$$

so that

$$(42) \qquad G_j(x) + x \int_{\{\lambda > x\}} \frac{dG_j(x)}{\lambda} = x \int_{\{\lambda > x\}} \frac{G_j(\lambda)}{\lambda^2}\, d\lambda,$$

and hence from (40),

$$(43) \qquad \int_{\{\lambda > x\}} \frac{G_1(\lambda)}{\lambda^2}\, d\lambda = \int_{\{\lambda > x\}} \frac{G_2(\lambda)}{\lambda^2}\, d\lambda$$

at every continuity point $x > 0$ of G_1 and G_2. Differentiating with respect to x gives $G_1(x) = G_2(x)$ for every such x and hence $G_1 = G_2$. (*Cf.* [13] and references for the question of when the "identifiability" assumption (D) holds.)

The defect of the method of Theorem 2 for estimating G is that the choice of G_n satisfying (9) is non-constructive. In contrast, the method of Section 4, which bypasses the estimation of G and gives t_G directly, is quite explicit for the given parametric family and loss structure.

In the next section we shall indicate a method for estimating G which works in the case of a parameter space Λ consisting of a *finite* number of elements, and which may possibly be capable of generalization.

7. Estimating the *a priori* distribution: the finite case.

Consider the case in which the observable random variable $x \varepsilon X$ is known to have one of a *finite* number of specified probability distributions $P_1, \cdots, P_r$, which one depending on the value of a random parameter $\lambda \varepsilon \Lambda = \{1, \cdots, r\}$ which has an unknown *a priori* probability vector $G = \{g_1, \cdots, g_r\}$, $g_i \geqq 0$, $\sum_{i=1}^r g_i = 1$, such that $P(\lambda = i) = g_i$. Observing a sequence $x_1, x_2, \cdots$ of independent random variables with the common distribution

$$P_G(x \varepsilon B) = \sum_1^r g_i P_i(B), \tag{1}$$

our problem is to construct functions

$$g_{i,n} = g_{i,n}(x_1, \cdots, x_n) \tag{2}$$

such that $g_{i,n} \geqq 0$, $\sum_{i=1}^r g_{i,n} = 1$, and whatever be G,

$$P[\lim_{n\to\infty} g_{i,n} = g_i] = 1 (i = 1, \cdots, r). \tag{3}$$

A *necessary* condition for the existence of such a sequence $g_{i,n}$ is clearly that

(A) If $G = \{g_1, \cdots, g_r\}$ and $\bar{G} = \{\bar{g}_1, \cdots, \bar{g}_r\}$

are any two probability vectors such that for every B,

$$\sum_{i=1}^r g_i P_i(B) = \sum_{i=1}^r \bar{g}_i P_i(B) \tag{4}$$

then $G = \bar{G}$. We shall now show that (A) is also *sufficient*.

Denote by μ any σ-finite measure on X with respect to which all the P_i are absolutely continuous and such that their densities $f_i = dP_i/d\mu$ are square integrable:

$$\int_X f_i^2(x)\, d\mu(x) < \infty \qquad (i = 1, \cdots, r). \tag{5}$$

(For example, we can always take $\mu = P_1 + \cdots + P_r$, since then $0 \leqq f_i(x) \leqq 1$ and hence

$$\int_X f_i^2(x)\, d\mu(x) \leqq \int_X f_i(x)\, d\mu(x) = 1.) \tag{6}$$

The functions f_i are elements of the Hilbert space H over the measure space (X, μ). From (A) they are *linearly independent*. For if $c_1 f_1 + \cdots + c_r f_r = 0$ for some constants c_i not all 0, then by renumbering the f_i we can write

$$c_1 f_1 + \cdots + c_k f_k = c_{k+1} f_{k+1} + \cdots + c_q f_q \tag{7}$$

with $c_1, \cdots, c_q$ all positive and $1 \leqq q \leqq r$. Integrating over X we obtain

$$c_1 + \cdots + c_k = c_{k+1} + \cdots + c_q = c > 0, \tag{8}$$

and hence

$$(9)\quad G = \{c_1/c, \cdots, c_k/c, 0, \cdots, 0\} \neq \bar{G} = \{0, \cdots, 0, c_{k+1}/c, \cdots, c_q/c\}$$

are such that (4) holds, contradicting (A).

Now let H_j denote the linear manifold spanned by the $r - 1$ functions $f_1, \cdots, f_{j-1}, f_{j+1}, \cdots, f_r$. We can then write uniquely

$$(10)\qquad f_j = f_j' + f_j'' \qquad (j = 1, \cdots, r)$$

with

$$(11)\qquad f_j' \varepsilon H_j, \qquad f_j'' \perp H_j, \qquad f_j'' \neq 0.$$

Hence, setting

$$(12)\qquad \phi_j(x) = f_j''(x) \Big/ \int_X [f_j''(x)]^2\, d\mu(x),$$

we have

$$(13)\qquad \int_X \phi_j(x) f_k(x)\, d\mu(x) = \begin{cases} 1 & \text{if } j = k, \\ 0 & \text{if } j \neq k. \end{cases}$$

Now define

$$(14)\qquad \bar{g}_{i,n} = n^{-1} \sum_{\nu=1}^{n} \phi_i(x_\nu), \quad g_{i,n} = [\bar{g}_{i,n}]^+ / \sum_{j=1}^{r} [\bar{g}_{j,n}]^+,$$

where $[a]^+$ denotes $\max(a, 0)$. If $x_1, x_2, \cdots$ are independent random variables with the common distribution (1), their common density with respect to μ is

$$(15)\qquad \sum_{j=1}^{r} g_j f_j(x),$$

so that by (13),

$$(16)\quad E\phi_i(x_\nu) = \int_X \phi_i(x) \sum_{j=1}^{r} g_j f_j(x)\, d\mu(x) = \sum_{j=1}^{r} g_j \int_X \phi_i(x) f_j(x)\, d\mu(x) = g_j.$$

The strong law of large numbers then implies (3).

Returning now to the problem of Section 2 in which we have an action space A and a loss structure $L(a, \lambda)$, here specified by functions $L(a, i)$ $(i = 1, \cdots, r)$, let us assume for simplicity that

$$(17)\qquad 0 \leqq L(a, i) \leqq L < \infty \quad (\text{all } a \varepsilon A \text{ and } i = 1, \cdots, r).$$

Then (2.6) becomes

$$(18)\qquad \Delta_G(a, x) = \sum_{i=1}^{r} [L(a, i) - L(a_0, i)] f_i(x) g_i,$$

and we can set

$$\Delta_n(a, x) = \sum_{i=1}^{r} [L(a, i) - L(a_0, i)]f_i(x)g_{i,n}, \tag{19}$$

so that

$$\sup_a |\Delta_n(a, x) - \Delta_G(a, x)| \leqq L\sum_{i=1}^{n} f_i(x)|g_i - g_{i,n}|. \tag{20}$$

Since $f_i(x) < \infty$ for a.e. $(\mu)x$ it follows from (3) that with probability 1, (2.10) holds, so that $T = \{t_n\}$ defined by (2.11) is a.o. relative to every $G = (g_1, \cdots, g_r\}$.

It would be interesting to try to extend this method of estimating G to the case of a continuous parameter space, say $\Lambda = \{-\infty < \lambda < \infty\}$. One possible way is the following. Suppose for definiteness that λ is the location parameter of a normal distribution with unit variance, so that $x_1, x_2, \cdots$ have the common density function

$$f_G(x) = \int_{\infty}^{\infty} f(x - \lambda)\, dG(\lambda); \qquad f(x) = (2\pi)^{-\frac{1}{2}}e^{-\frac{1}{2}x^2} \tag{21}$$

with respect to Lebesgue measure on $X = \{-\infty < x < \infty\}$.

For any $n \geqq 1$ let

$$\lambda_1^{(n)} < \lambda_2^{(n)} < \cdots < \lambda_{k_n}^{(n)} \tag{22}$$

be constants, and let $g_{i,n}$ $(i = 1, \cdots, k_n)$ be defined by (14) where the $f_j(x)$ of (5) are replaced by $f(x - \lambda_j^{(n)})$. Consider the random distribution function

$$G_n(\lambda) = \sum g_{i,n} \quad \text{(sum over all } i \text{ such that } \lambda_i^{(n)} \leqq \lambda). \tag{23}$$

Can we choose the values k_n and (22) for each n so that, whatever be G,

$$P[G_n \rightarrow G] = 1? \tag{24}$$

REFERENCES

[1] HANNAN, J. F. and ROBBINS, H. (1955). Asymptotic solutions of the compound decision problem for two completely specified distributions. *Ann. Math. Statist.* **26** 37–51.

[2] JOHNS, M. V., JR. (1956). Contributions to the theory of non-parametric empirical Bayes procedures in statistics. Columbia Univ. Dissertation.

[3] JOHNS, M. V., JR. (1957). Non-parametric empirical Bayes procedures. *Ann. Math. Statist.* **28** 649–669.

[4] JOHNS, M. V., JR. (1961). An empirical Bayes approach to non-parametric two-way classification. *Studies in Item Analysis and Prediction* (ed. by H. Solomon), pp. 221–232. Stanford Univ. Press.

[5] MIYASAWA, K. (1961). An empirical Bayes estimator of the mean of a normal population. *Bull. Inst. Internat. Statist.* **38** 181–188.

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

[7] ROBBINS, H. (1950). Asymptotically subminimax solutions of compound statistical decision problems. *Proc. Second Berkeley Symp. Math. Statist. Prob.* 131–148.

[8] ROBBINS, H. (1955). An empirical Bayes approach to statistics. *Proc. Third Berkeley Symp. Math. Statist. Prob.* **1** 157–164.

[9] ROBBINS, H. (1963). The empirical Bayes approach to testing statistical hypotheses. To appear in *Rev. Inst. Internat. Statist.*

[10] SAMUEL, E. (1963). Asymptotic solutions of the sequential compound decision problem. *Ann. Math. Statist.* **34** 1079–1094.

[11] SAMUEL, E. (1963). An empirical Bayes approach to the testing of certain parametric hypotheses. *Ann. Math. Statist.* **34** 1370–1385.

[12] SAMUEL, E. Strong convergence of the losses of certain decision rules for the compound decision problem. Unpublished.

[13] TEICHER, H. (1961). Identifiability of mixtures. *Ann. Math. Statist.* **32** 244–248.