AN EMPIRICAL BAYES APPROACH TO THE TESTING OF CERTAIN PARAMETRIC HYPOTHESES

By Ester Samuel

The Hebrew University of Jerusalem

1. Summary. The empirical Bayes approach is described in Section 2. In Section 3 "optimal" empirical Bayes rules are given for the problem of testing a simple hypothesis against a simple alternative. In Section 4 a limit theorem is proved which is used in Section 5 to obtain "optimal" empirical Bayes rules for testing one and two-sided hypotheses about the parameters in the Poisson, geometric, negative binomial and binomial distributions. The same methods are used in Section 6 to obtain "optimal" empirical Bayes rules for testing hypotheses about parameters in continuous distributions of the exponential family. Examples of areas of applications are given in Section 7, and the last Section discusses uses of the above methods in the compound decision problem.

2. Introduction. Consider a random variable X distributed according to distribution function $P(x \mid \lambda) = P(X \leqq x \mid \lambda)$ where λ is a real parameter known to belong to some set Ω, and for each $\lambda \varepsilon \Omega$ $P(x \mid \lambda)$ is completely specified. A statistical decision problem arises if the statistician has to take some action, and the best action depends on λ, but λ is unknown. In particular, for each action A in a set of possible actions $\mathfrak{A}$, there is defined a (usually non-negative) loss function $L = L(A, \lambda)$. Since λ is unknown the statistician will base his decision on the observation x on X. Let $\varphi(x)$ denote a (randomized) decision function, i.e. for each observed value x of X, $\varphi(x)$ is a distribution over the set $\mathfrak{A}$ of possible actions. The risk function corresponding to φ, i.e. the expected loss, is

$$R(\varphi, \lambda) = E_\lambda \int_{\mathfrak{A}} L(A, \lambda)\, d\varphi \tag{1}$$

where E_λ denotes the expectation with respect to distribution $P(x \mid \lambda)$. The statistician's aim is thus to minimize (1), but, as is well known, except in trivial cases there exists no rule φ minimizing (1) uniformly in $\lambda \varepsilon \Omega$.

In the Bayesian approach the parameter λ is considered a realization of a random variable Λ, distributed according to some distribution function G on Ω, where $G(\lambda) = P(\Lambda \leqq \lambda)$. In this case it is natural to consider the "global" expected loss, or Bayes risk, of φ, rather than (1). It is defined by $R(\varphi, G) = \int_\Omega R(\varphi, \lambda)\, dG$. For given G there will usually exist a Bayes rule with respect to apriori distribution G, denoted φ_G, for which $\min_\varphi R(\varphi, G) = R(\varphi_G, G) = R(G)$. $R(G)$ is called the "Bayes envelope function", and denotes the minimal possible global risk attainable by any decision function if λ is the realization of a random variable with distribution G. Though the assumption of the existence of an apriori distribution G is often reasonable, it is usually difficult in

Received February 11, 1963; revised June 13, 1963.

Institute of Mathematical Statistics is collaborating with JSTOR to digitize, preserve, and extend access to *The Annals of Mathematical Statistics.* www.jstor.org

practice to assert what this distribution is, and thus one will usually use a rule the Bayes risk of which exceeds $R(G)$.

Consider now the situation where $(X_1, \Lambda_1), (X_2, \Lambda_2), \cdots, (X_n, \Lambda_n), \cdots$ is a sequence of independent pairs of random variables, where the Λ_n's are identically distributed according to G, and where for $n = 1, 2, \cdots$, $P(X_n \leqq x \mid \Lambda_n = \lambda) = P(x \mid \lambda)$. We assume that for each n, $n = 1, 2, \cdots$ the statistician is confronted with the same decision problem, and that these problems arise sequentially. Even if G is unknown, we may hope that as n increases it will be possible to improve the decision rule used, since at the time when the statistician must decide on λ_n he has all the observations $x_1, x_2, \cdots, x_n = \mathbf{x}_n$ on $X_1, X_2, \cdots, X_n = \mathbf{X}_n$ at his disposal, and the distribution of every X is

$$P_G(x) = P(X \leqq x) = \int_\Omega P(x \mid \lambda)\, dG. \tag{2}$$

A method using previous observations in order to approach the Bayes decision rule φ_G was first used by Robbins in [12], where it is called an "empirical Bayes approach". The decision problem considered in [12] is one of estimating λ for various families of distributions $P(x \mid \lambda)$. Johns in [6] and [7] generalizes the results of [12] and shows that for some of the proposed procedures not only the rules, but also their risks converge to the corresponding $R(G)$, whatever be G. In [8] Johns considers an empirical Bayes problem related to the subject matter of the present paper.

In the present paper we discuss the empirical Bayes approach to the two-decision problem, focussing particularly on the problem of testing one sided hypotheses of the kind

$$\begin{matrix} H_0 : \lambda \leqq \lambda^* \\ H_1 : \lambda > \lambda^* \end{matrix} \quad \text{and (3a)} \quad \begin{matrix} H_0 : \lambda \geqq \lambda^* \\ H_1 : \lambda < \lambda^* \end{matrix} \tag{3}$$

and two sided hypotheses of the kind

$$\begin{matrix} H_0 : |\lambda - \lambda^*| \leqq \Delta \\ H_1 : |\lambda - \lambda^*| > \Delta \end{matrix} \quad \text{and (4a)} \quad \begin{matrix} H_0 : |\lambda - \lambda^*| \geqq \Delta \\ H_1 : |\lambda - \lambda^*| < \Delta \end{matrix} \tag{4}$$

where λ^* and $\Delta > 0$ are fixed constants. Since we shall consider only two-decision problems, we shall denote the two actions available by A_0 and A_1, where A_i can usually be interpreted to mean: "say H_i is correct", $i = 0, 1$. For such problems any decision rule φ can be written simply as a (measurable) function $t = t(x)$ where $0 \leqq t \leqq 1$ and $t(x)$ and $1 - t(x)$ denote the probabilities of deciding A_1 and A_0 respectively, once $X = x$ is observed. In the empirical Bayes situation, where we have $\mathbf{x}_n$ at our disposal at the nth decision, we shall use $t_n = t_n(\mathbf{x}_n)$.

In the next section we consider an empirical Bayes solution of the general simple versus simple hypothesis testing problem, and in Section 5 the problem of testing (3) and (4) for discrete distributions of the type

$$P(X = x \mid \lambda) = \lambda^x h(\lambda) g(x) \quad \text{for} \quad x = 0, 1, \cdots \tag{5}$$

is considered. The Poisson, geometric and general negative binomial distributions are particular cases of (5). Section 6 considers the problem of testing (3) and (4) for random variables with densities (with respect to Lebesgue measure) belonging to the exponential family.

3. The empirical Bayes approach to testing a simple hypothesis versus a simple alternative. The simple versus simple hypothesis testing problem can be stated as $H_0 : P = P_0$, $H_1 : P = P_1$ where P is the distribution function of a random variable X and where P_0 and P_1 are completely specified distributions, with densities $f(x \mid 0)$ and $f(x \mid 1)$ respectively, with respect to some measure μ. Here it is natural to consider the loss function

$$\begin{aligned} L(A_j, \lambda) &= 0 \quad j = \lambda \quad \lambda = 0, 1 \\ &= a \quad j = 0 \quad \lambda = 1 \\ &= b \quad j = 1 \quad \lambda = 0 \end{aligned}$$

where $a > 0, b > 0$.

In this case Λ is a Bernoulli random variable and G corresponds to a Bernoulli apriori distribution. Let $h(x)$ be a bounded unbiassed estimate of λ and let $p_n = p_n(\mathbf{x}_n)$ equal 0, $\sum_{j=1}^n h(x_j)/n$ and 1 as $\sum_{j=1}^n h(x_j)/n$ is less than zero, between zero and one, and exceeds one respectively. Then it follows easily from the results of Hannan and Robbins in [5] that the rule with

$$\begin{aligned} t_n(\mathbf{x}_n) &= 1 \quad \text{if} \quad p_n a f(x_n \mid 1) > (1 - p_n) b f(x_n \mid 0) \\ &= 0 \qquad \text{otherwise} \end{aligned} \tag{6}$$

satisfies

$$\lim_{n\to\infty} R(t_n, G) = R(G) \tag{7}$$

whatever be the (Bernoulli) apriori distribution G. Other (related) rules satisfying (7) are given in [15]. See also [16]. A direct proof that (6) satisfies (7) can also be found in Robbins [13], a manuscript-copy of which was obtained by the author after the present paper was submitted. [13] is also closely related to the other subject matter of the present paper.

4. A theorem. Let $L(A_i, \lambda)$ denote the loss function when action A_i is taken, $i = 0, 1$. Then for any decision function $t = t(x)$ the Bayes risk function with respect to apriori distribution G is given by

$$\begin{aligned} R(t, G) &= E[t(X)L(A_1, \Lambda) + (1 - t(X))L(A_0, \Lambda)] \\ &= E_G[L(A_0, \Lambda)] - E[t(X)(L(A_0, \Lambda) - L(A_1, \Lambda))] \\ &= E_G[L(A_0, \Lambda)] - E[t(X)E_G[L(A_0, \Lambda) - L(A_1, \Lambda) \mid X]], \end{aligned} \tag{8}$$

where E and E_G denote the expectations with respect to the joint distribution of (X, Λ) and of Λ respectively, and where $E_G[\cdot \mid X]$ denotes the conditional

expectation of Λ given X. Since the first term in the right hand side of (8) does not involve X, it follows that the t's which minimize (8), for fixed G, are

$$
\begin{aligned}
t_G(x) &= 1 && \text{as } E_G[L(A_0, \Lambda) - L(A_1, \Lambda) \mid x] > 0 \\
&= 0 && \qquad\qquad\quad < 0 \\
&= \text{arbitrary in } [0, 1] && \qquad\qquad\quad = 0.
\end{aligned} \tag{9}
$$

In order to avoid randomization we shall consider the version of $t_G(x)$ where the arbitrary part is taken to be 0. This version is of the form

$$
\begin{aligned}
t^K(x) &= 1 \quad \text{if} \quad K(x) > 0 \\
&= 0 \quad \text{if} \quad K(x) \leqq 0.
\end{aligned} \tag{10}
$$

For any function $K(x)$ it follows from (8) that the Bayes risk of t^K is

$$
\begin{aligned}
R(t^K, G) = E_G[\mathrm{L}(A_0, \Lambda)] &- E_G[L(A_0, \Lambda) \\
&- L(A_1, \Lambda) \mid K(X) > 0]P[K(X) > 0].
\end{aligned} \tag{11}
$$

Let

$$
\begin{aligned}
t_*^K(x) &= 1 && \text{if} \quad K(x) > 0 \\
&= 0 && \text{if} \quad K(x) < 0 \\
&= \text{arbitrary in } [0, 1] && \text{if} \quad K(x) = 0.
\end{aligned}
$$

Suppose $K(x)$ is such that

$$
R(t^K, G) = R(t_*^K, G) \quad \text{for all} \quad t_*^K. \tag{12}
$$

(This is satisfied in particular for t_G defined in (9).)

We shall now consider the case where instead of a fixed function $K(x)$ we have a sequence of random functions $K_n(x; \mathbf{X}_n)$, where $\mathbf{X}_n$ is a random vector independent of (X, Λ). Let $t_n^K = t_n^K(x; \mathbf{X}_n)$ be the (random) decision function defined for $K_n(x; \mathbf{X}_n)$ by (10), i.e. decide

$$
\begin{aligned}
t_n^K(x; \mathbf{X}_n) &= 1 \quad \text{if} \quad K_n(x; \mathbf{X}_n) > 0 \\
&= 0 \quad \text{if} \quad K_n(x; \mathbf{X}_n) \leqq 0
\end{aligned} \tag{13}
$$

when $X = x$ is observed.

The Bayes risk $R(t_n^K, G)$ will be an expression corresponding to (8) (or (11)) where the expectation E (and probability P in (11)) is now taken over the joint distribution of (X, Λ) and $\mathbf{X}_n$. We have the following

THEOREM. *Let $K_n(x; \mathbf{X}_n)$ be such that for each x*

$$
K_n(x; \mathbf{X}_n) \to K(x) \quad \textit{in probability} \tag{14}
$$

(where "in probability" refers to the distribution of $\mathbf{X}_n$) and suppose (12) holds for $K(x)$.

If

$$E_G[|L(A_0, \Lambda) - L(A_1, \Lambda)|] < \infty \tag{15}$$

then the rule t_n^K *with*

$$\begin{aligned} t_n^K(x; \mathbf{x}_n) &= 1 \quad \text{if} \quad K_n(x; \mathbf{x}_n) > 0 \\ &= 0 \quad \text{if} \quad K_n(x; \mathbf{x}_n) \leqq 0 \end{aligned} \tag{16}$$

has a risk satisfying

$$\lim_{n\to\infty} R(t_n^K, G) = R(t^K, G). \tag{17}$$

In particular, if

$$K(x) = E_G[L(A_0, \Lambda) - L(A_1, \Lambda)| x] \tag{18}$$

then (12) *holds and* (17) *becomes* $\lim_{n\to\infty} R(t_n^K, G) = R(G)$.

PROOF. From (8) it follows that it suffices to show that

$$\begin{aligned} &\lim_{n\to\infty} E[t_n^K(X; \mathbf{X}_n) E_G[L(A_0, \Lambda) - L(A_1, \Lambda)| X] \\ &\qquad = E[t^K(X) E_G[L(A_0, \Lambda) - L(A_1, \Lambda) | X] \end{aligned} \tag{19}$$

where E denotes the expectation with respect to the joint distribution of (X, Λ) and $\mathbf{X}_n$. $\mathbf{X}_n$ is assumed to be independent of X and Λ, and we shall denote the expectation with respect to $\mathbf{X}_n$ by E_n, and with respect to (X, Λ) by E_*. Now it follows by (13) and (14) that for each x, except possibly for values of x for which $K(x) = 0$

$$\lim_{n\to\infty} E_n[t_n^K(x; \mathbf{X}_n)] = t^K(x). \tag{20}$$

Also

$$\begin{aligned} &E[t_n^K(X; \mathbf{X}_n) E_G[L(A_0, \Lambda) - L(A_1, \Lambda)| X]] \\ &\qquad = E_*(E_n[t_n^K(X; \mathbf{X}_n)| X] E_G[L(A_0, \Lambda) - L(A_1, \Lambda)| X]). \end{aligned} \tag{21}$$

Now

$$\begin{aligned} &|E_n[t_n^K(X; \mathbf{X}_n)|X] E_G[L(A_0, \Lambda) - L(A_1, \Lambda)| X]| \\ &\qquad \leqq E_G[|\, L(A_0, \Lambda) - L(A_1, \Lambda)|| X] \end{aligned}$$

and by (15)

$$\begin{aligned} &E(E_G[|\, L(A_0, \Lambda) - L(A_1, \Lambda)|| X]) \\ &\qquad = E_G[|L(A_0, \Lambda) - L(A_i, \Lambda)|] < \infty. \end{aligned}$$

Thus (19) follows from (21) by the use of (20), (12) and Lebesgue's dominated convergence theorem.

5. Application of the Theorem for two action problems involving the Poisson, geometric, negative binomial and binomial distributions. In applications of the Theorem we shall try to look for a sequence $K_n(x; \mathbf{X}_n)$ of functions for which (14) holds with $K(x)$ defined in (18), since this will minimize the possible limit

of $R(t_n^K, G)$. Since $K(x)$ in (18) is a function of the (unknown) apriori distribution G, we would like (14) to hold for (18) for all possible distributions G. Obviously this may be possible only if the distribution of $\mathbf{X}_n$ has some relation to G. We shall see that in the empirical Bayes approach the limit $R(G)$ can sometimes be attained.

We shall let $\mathbf{X}_n$ denote the vector of random variables $X_1, \cdots, X_n$ on which the first n observations are taken, and we assume that X_i are independent, each with distribution function (2). For X we shall take X_{n+1} which again is assumed independent, with distribution (2).

Consider the case where $P(x \mid \lambda)$ is an integer valued discrete distribution and let $p(x \mid \lambda) = P(X = x \mid \Lambda = \lambda)$ $x = 0, 1, \cdots$, and $p_G(x) = P_G(X = x) = \int_\Omega p(x \mid \lambda)\, dG(\lambda)$, $x = 0, 1, \cdots$. $p_G(x)$ is the unconditional distribution of X_i when Λ is distributed according to G.

Define $p_n(x) = p_n(x; \mathbf{X}_n) =$ (number of indices i, $i = 1, \cdots, n$, for which $X_i = x)/n$, $x = 0, 1, \cdots$. $p_n(x)$ is the empirical probability function of $\mathbf{X}_n$ and, as is well known

$$\lim_{n\to\infty} p_n(x) = P_G(x) \tag{22}$$

for all x, with probability one, whatever be the (unknown) apriori distribution G.

We shall consider the case where

$$L(A_0, \lambda) - L(A_1, \lambda) = \sum_{j=0}^{s} a_j \lambda^j, \tag{23}$$

i.e. the difference between the losses is a polynomial in λ. For (23), (15) holds whenever G has a finite moment of order s. Particular cases of (23) are

$$\begin{aligned} L(A_0, \lambda) &= c(\lambda - \lambda^*) && \text{for} && \lambda > \lambda^* \\ &= 0 && \text{otherwise} \\ L(A_1, \lambda) &= c(\lambda^* - \lambda) && \text{for} && \lambda < \lambda^* \\ &= 0 && \text{otherwise} \end{aligned} \tag{24}$$

and

$$\begin{aligned} L(A_0, \lambda) &= 0 && \text{if} \quad |\lambda - \lambda^*| \leqq \Delta \\ &= c[(\lambda - \lambda^*)^2 - \Delta^2] && \text{otherwise} \\ L(A_1, \lambda) &= 0 && \text{if} \quad |\lambda - \lambda^*| > \Delta \\ &= c[\Delta^2 - (\lambda - \lambda^*)^2] && \text{otherwise} \end{aligned} \tag{25}$$

where $c > 0$ is some fixed constant. (24) and (25) seem particularly proper for testing (3) and (4) respectively. They may, in fact, be more reasonable than the usual zero-one losses. (24) and (25) are considered also in [7]. For (23) one has

$$E_G[L(A_0, \Lambda) - L(A_1, \Lambda) \mid x] = \frac{\int_\Omega \left(\sum_{j=0}^{s} a_j \lambda^j\right) p(x \mid \lambda)\, dG(\lambda)}{\int_\Omega p(x \mid \lambda)\, dG(\lambda)}. \tag{26}$$

For distributions of type (5)

$$p_G(x) = g(x) \int_\Omega \lambda^x h(\lambda)\, dG(\lambda) \tag{27}$$

and (26) becomes

$$E_G[L(A_0, \Lambda) - L(A_1, \Lambda) \mid x] = \frac{\sum_{j=0}^{s} a_j \int_\Omega \lambda^{x+j} h(\lambda)\, dG(\lambda)}{\int_\Omega \lambda^x h(\lambda)\, dG(\lambda)} = \frac{\sum_{j=0}^{s} a_j\, p_G(x+j)/g(x+j)}{p_G(x)/g(x)}. \tag{28}$$

Thus it follows from (22) that

$$K_n(x; \mathbf{X}_n) = \frac{\sum_{j=0}^{s} a_j\, p_n(x+j)/g(x+j)}{p_n(x)/g(x)} \tag{29}$$

converges to (28) with probability one, and the Theorem is applicable and (16) yields an "optimal" empirical Bayes rule, provided (15) holds.

Thus for distributions of type (5) the empirical Bayes rule with

$$t_{n+1} = t_{n+1}(\mathbf{x}_{n+1})$$

where

$$\begin{aligned} t_{n+1}(\mathbf{x}_{n+1}) &= 1 \quad \text{if} \quad K_n(x_{n+1}\,; \mathbf{x}_n) > 0 \\ &= 0 \qquad \text{otherwise} \end{aligned}$$

satisfies $\lim_{n\to\infty} R(t_n, G) = R(G)$ whatever be G, provided (15) holds.

Particular distributions of type (5) are the Poisson, geometric and negative binomial distributions, (the latter depending on a positive integer-valued nuisance parameter m), where for

Poisson	$g(x) = 1/x!$
geometric	$g(x) = 1$
negative binomial	$g(x) = \binom{x+m-1}{x} \quad x = 0, 1, \cdots$.

Thus, for example, for testing (3) with loss function (24) $K_n(x; \mathbf{X}_n)$ becomes

$$\begin{aligned} K_n(x; \mathbf{X}_n) &= c[(x+1)p_n(x+1)/p_n(x) - \lambda^*] && \text{for the Poisson} \\ &= c[p_n(x+1)/p_n(x) - \lambda^*] && \text{for the geometric} \\ &= c[(x+1)p_n(x+1)/(x+m)p_n(x) - \lambda^*] && \text{for the negative binomial} \end{aligned}$$

distribution, and an optimal empirical Bayes rule $t_{n+1} = t_{n+1}(\mathbf{x}_{n+1})$ becomes: On the $(n+1)$st decision take action A_1 when $\mathbf{X}_{n+1} = \mathbf{x}_{n+1}$ is observed if and only if

$$(30)\quad \begin{cases} (x_{n+1}+1)p_n(x_{n+1}+1)/p_n(x_{n+1}) > \lambda^* & \text{for the Poisson} \\ p_n(x_{n+1}+1)/p_n(x_{n+1}) > \lambda^* & \text{for the geometric} \\ (x_{n+1}+1)p_n(x_{n+1}+1)/(x_{n+1}+m)p_n(x_{n+1}) > \lambda^* & \text{for the negative binomial} \end{cases}$$

distribution. For the geometric and negative binomial distributions $0 \leqq \lambda \leqq 1$, so (15) is no restriction for losses of type (23). The left hand sides of (30) are easily computable, but the statistician may prefer to start using t_n only for n sufficiently large, since for small n, $p_n(x)$ will often be 0, which may cause some trouble. It is interesting to notice that the left hand sides of (30) provide optimal empirical Bayes solutions to the corresponding problems of estimating λ when G is unknown, where the loss it taken to be the squared deviation. For the Poisson and geometric distributions these estimates are given already in [12].

For testing (3a) an appropriate loss function is one where $L(A_i, \lambda)$ is $L(A_{|i-j|}, \lambda)$ of (24) $i, j = 0, 1$ $i \neq j$, and (30) with reversed inequalities is an optimal empirical Bayes rule. Likewise, optimal empirical Bayes rules for these distributions, for testing (4) with loss (25) are easily obtainable by substituting from (25) in (29). Case (4a) is treated similarly.

Consider now the *binomial distribution*, i.e.

$$P(X = x) = p_r(x \mid \lambda) = \binom{r}{x}\lambda^x(1-\lambda)^{r-x} \qquad x = 0, 1, \cdots, r.$$

For this distribution

$$(31)\qquad p_{G,r}(x) = \binom{r}{x}\int_0^1 \lambda^x(1-\lambda)^{r-x}\,dG(\lambda) \qquad x = 0, 1, \cdots, r$$

and for losses satisfying (23) we have

$$(32)\qquad \begin{aligned} E_{G,r}[L(A_0, \Lambda) - L(A_1, \Lambda) \mid x] &= \frac{\sum_{j=0}^{s} a_j \int_0^1 \lambda^{x+j}(1-\lambda)^{r-x}\,dG(\lambda)}{\int_0^1 \lambda^x(1-\lambda)^{r-x}\,dG(\lambda)} \\ &= \sum_{j=0}^{s}\left[a_j\, p_{G,r+j}(x+j) \Big/ \binom{r+j}{x+j}\right] \Big/ \left[p_{G,r}(x) \Big/ \binom{r}{x}\right]. \end{aligned}$$

Since the right hand side of (32) involves the function $p_{G,r+s}$ and we are dealing with binomial random variables with parameter r only, we cannot hope to find a function which will converge to (32), but we can still use the theorem to obtain "good" empirical Bayes rules for this situation.

Considering the binomial variable X_i as the number of successes in r independent Bernoulli trials with probability Λ_i of success, we assume that not only the total, X_i, but also the order of successes is known. Let $X_i^{(v)}$ denote the number of successes in the v first of the r trials of X_i, $v = 1, 2, \cdots, r$, and define the empirical distribution functions

$$p_n^{(v)}(x) = p_n^{(v)}(x; \mathbf{X}_n^{(v)})$$

$$= (1/n) \text{ (number of indices } i, i = 1, \cdots, n \text{ for which } X_i^{(v)} = x), \quad x = 0, 1, \cdots, v.$$

Since $\lim_{n\to\infty} p_n^{(v)}(x) = p_{G,v}(x)$ for $x = 0, \cdots, v$ with probability one, it follows from (32) that (for $r > s$)

$$K_n(x; \mathbf{X}_n) = \frac{\sum_{j=0}^{s} a_j \, p_n^{(r-s+j)}(x+j) \Big/ \binom{r-s+j}{x+j}}{p_n^{(r-s)}(x) \Big/ \binom{r-s}{x}}$$

converges to

$$E_{G,r-s}[L(A_0, \Lambda) - L(A_1, \Lambda) \mid x] \qquad \text{for} \quad x = 0, 1, \cdots, r-s \tag{33}$$

with probability one. Since (12) holds for (33) it follows by the theorem that if we let

$$\begin{aligned} t_{n+1}(\mathbf{x}_{n+1}) &= 1 \quad \text{if} \quad K_n(x_{n+1}^{(r-s)}; \mathbf{x}_n) > 0 \\ &= 0 \qquad \text{otherwise} \end{aligned} \tag{34}$$

then

$$\lim_{n\to\infty} R(t_n, G) = R_{r-s}(G), \quad \text{for all} \quad G \tag{35}$$

where $R_v(G)$ denotes the Bayes envelope function for the above discussed loss function and a binomial distribution with parameter v. Thus, with the above described rule one will in the limit be just as well off as if one knew G but based the decision on a binomial variable with parameter $r - s$ rather than r. The difference $R_{r-s}(G) - R_r(G)$ may be considered the "cost of ignorance of G".

As a particular example we shall again consider problem (3) and loss (24). Here (34) becomes: Take action A_1 if and only if

$$x_{n+1}^{(r-1)} p_n^{(r)}(x_{n+1}^{(r-1)} + 1)/r p_n^{(r-1)}(x_{n+1}^{(r-1)}) > \lambda^*. \tag{36}$$

(The left hand side of (36) is again an estimate of λ given in [12], and the present method of dealing with the binomial case is motivated by Robbins' work [12].) It follows from (35) that the risk of this rule converges to $R_{r-1}(G)$, whatever be G. The difference $R_{r-1}(G) - R_r(G)$ is thus of interest. Consider (24) with $c = 1$ and $\lambda^* = \frac{1}{2}$. If G is degenerate, i.e. assigns probability one to some value λ_0, then

$R_{r-1}(G) = R_r(G) = 0$. The values of $R_r(G)$ in the case where G is the uniform distribution over the interval [0, 1] are given in the table below.

r	$R_r(G)$
1, 2	$\frac{1}{24} = .042$
3, 4	$\frac{1}{40} = .025$
5, 6	$\frac{1}{56} = .018$
7, 8	$\frac{1}{72} = .014$

Thus for this example $R_{r-1}(G) - R_r(G) > 0$ if and only if r is an odd integer, and this difference decreases rapidly with increasing r.

One may wonder why the binomial distribution does not give rise to as good empirical Bayes rules as the three other distributions discussed. The reason, pointed out to the author in a discussion with Professor Robbins, can be found when comparing (31) and (27). In the empirical Bayes approach we use $\mathbf{X}_n$ in order to obtain "information" about G and this "information" is contained in an estimate of $p_G(x)$ which converges to $p_G(x)$ with probability 1. Now for the binomial distribution it is seen from (31) that for $x = 0, 1, \cdots, r$ $p_{G,r}(x)$ is a linear combination of the r first moments of G. Thus all distributions G over [0, 1] which have indentical r first moments give rise to the same $p_{G,r}(x)$. However, for losses satisfying (23) it follows from (32) that the best procedure is a function of the $r + s$ first moments of G, and in order to obtain a limiting loss $R(G)$ an estimate of this function would be required, but such an estimate *cannot* be obtained from $\mathbf{X}_n$. Thus the exhibited rules (34) which are shown to satisfy (35) for all G must be considered "optimal in the limit". For the other three distributions, $p_G(x)$ given in (27) is a function of G depending on G in a more involved manner, and the mapping (2) of G to P_G is one-to-one when the class G is properly restricted. Thus in this case the "information" about G contained in $\mathbf{X}_n$ is much larger. (Compare [12] p. 162.)

We have exhibited empirical Bayes solutions for the four families of distributions considered above only in the case where the loss function satisfies (23). Since many functions can be approximated to any required degree of accuracy by a polynomial, the solution is quite general. It should, however, be remembered that in the binomial case our limiting risk will be $R_{r-s}(G)$ and not $R_r(G)$, when the degree of the polynomial is s, $(s < r)$.

6. Applications of the theorem for continuous distributions. In the previous section we considered applications of the theorem in various decision problems for several discrete distributions. Here we shall use a similar approach for the case where the independent, identically distributed random variables have a distribution function which is absolutely continuous with respect to the Lebesgue measure.

Let $f(x \mid \lambda)$ be the conditional density of X given $\Lambda = \lambda$, and for simplicity assume that the range for which $f(x \mid \lambda) > 0$ does not depend on λ. Then it

follows from Fubini's theorem that if Λ is distributed according to G the unconditional density of X (with respect to Lebesgue measure) is

$$f_G(x) = \int_\Omega f(x \mid \lambda)\, dG(\lambda). \tag{37}$$

In the empirical Bayesian situation $\mathbf{X}_n$ therefore constitutes a random sample of size n from a distribution with density (37). It would therefore be natural to seek an estimate $f_n(x) = f_n(x; \mathbf{X}_n)$, such that for all x, as $n \to \infty$

$$f_n(x) \to f_G(x) \quad \text{in probability} \tag{38}$$

for all possible f_G. $f_n(x)$ satisfying (38) is a consistent estimate of a density function. Such consistent estimates exist, and are exhibited by Rosenblatt [14] and Parzen [10]. One of the simplest classes of estimates satisfying (38), with some optimality properties (see [14]) is

$$f_n(x) = [F_n(x + h_n) - F_n(x - h_n)]/2h_n \tag{39}$$

where $h_n = dn^{-\frac{1}{5}}$, $d > 0$ some constant, and where $F_n(x) = F_n(x; \mathbf{X}_n)$ is the empirical distribution function defined by

$$\begin{aligned} &F_n(x) \\ &= (1/n)\ (\text{number of indices } i,\ i = 1, \cdots, n, \text{ for which } X_i \leqq x). \end{aligned} \tag{40}$$

For loss functions satisfying (23) one has

$$E_G[L(A_0, \Lambda) - L(A_1, \Lambda) \mid x] = \frac{\int_\Omega \left(\sum_{j=0}^{s} a_j \lambda^j\right) f(x \mid \lambda)\, dG(\lambda)}{\int_\Omega f(x \mid \lambda)\, dG(\lambda)}. \tag{41}$$

Corresponding to (5) we shall consider density functions of the form

$$\begin{aligned} f(x \mid \lambda) &= \lambda^x g(x) h(\lambda) \quad \text{for} \quad a < x < \infty \\ &= 0 \qquad\qquad\quad \text{otherwise,} \end{aligned} \tag{42}$$

where a is some constant (possibly $a = -\infty$). It then follows easily from the above that

$$K_n(x; \mathbf{X}_n) = \frac{\sum_{j=0}^{s} a_j f_n(x + j)/g(x + j)}{f_n(x)/g(x)}$$

converges in probability to (41) with $f(x \mid \lambda)$ given in (42). The Theorem therefore states that if on the $(n + 1)$st decision one takes action A_1 if and only if $K_n(x_{n+1}; \mathbf{x}_n) > 0$ when $\mathbf{X}_{n+1} = \mathbf{x}_{n+1}$ is observed, then the corresponding Bayes risk of the rules converges to $R(G)$ as $n \to \infty$, whatever be G, provided (15) holds.

(42) may seem a quite particular density function. The fact is, however,

that many well known density functions can be described by (42) after simple transformations. The following are some examples:

1. The *exponential distribution*: $0 < \theta < \infty$

$$f(x \mid \theta) = \theta e^{-\theta x} \quad \text{for} \quad x > 0$$
$$= 0 \qquad \text{otherwise.}$$

Set $e^{-\theta} = \lambda$. Then for $0 < \lambda < 1$

$$f(x \mid \lambda) = -\lambda^x \log \lambda \quad \text{for} \quad x > 0$$
$$= 0 \qquad \text{otherwise.}$$

2. The *normal distribution with fixed* σ: $-\infty < \theta < \infty$

$$f(x \mid \theta) = (2\pi\sigma^2)^{-\frac{1}{2}} \exp\left(-(x-\theta)^2/2\sigma^2\right) \quad \text{for} \quad -\infty < x < \infty.$$

Set $e^{\theta/\sigma^2} = \lambda$. Then for $0 < \lambda < \infty$

$$f(x \mid \lambda) = \lambda^x \exp\left(-x^2/2\sigma^2\right)\lambda^{-\sigma^2(\log\lambda)/2} \quad \text{for} \quad -\infty < x < \infty.$$

3. The *normal distribution with fixed* μ: $0 < \theta < \infty$

$$f(x \mid \theta) = (2\pi\theta^2)^{-\frac{1}{2}} \exp\left(-(x-\mu)^2/2\theta^2\right) \quad \text{for} \quad -\infty < x < \infty.$$

Set $y = (x-\mu)^2$, $\lambda = \exp(-1/2\theta^2)$. Then for $0 < \lambda < 1$

$$f(y \mid \lambda) = \lambda^y y^{-\frac{1}{2}}[(-\log\lambda)/\pi]^{\frac{1}{2}}/2 \quad \text{for} \quad y > 0$$
$$= 0 \qquad \text{otherwise.}$$

4. The *gamma distribution with* $p > 0$ *fixed*: $0 < \theta < \infty$

$$f(x \mid \theta) = [(2\theta^2)^{p/2}\Gamma(p/2)]^{-1}x^{p/2-1}\exp(-x/2\theta^2) \quad \text{for} \quad x > 0$$
$$= 0 \qquad \text{otherwise.}$$

Set $\lambda = \exp(-1/2\theta^2)$. Then for $0 < \lambda < 1$

$$f(x \mid \lambda) = \lambda^x x^{p/2-1}(-\log\lambda)^{p/2}/\Gamma(p/2) \quad \text{for} \quad x > 0$$
$$= 0 \qquad \text{otherwise.}$$

(Example 3 is a particular case of the latter, with $p = 1$.).

Many more examples can be given. The characteristic of all these examples is that they belong to the exponential family of distributions.

In all the above examples the function $\lambda = \psi(\theta)$ which takes the parameter θ into λ is a strictly monotone function of θ (for the range of θ for which $f(x \mid \theta)$ is a density function). Thus there is a one-to-one correspondence between hypotheses stated in terms of θ and those stated in terms of λ. If the original loss function (which is given as a function of θ) can be written in the form (23) as a function of λ and if the apriori distribution $G^*(\theta)$ corresponds to the apriori distribution $G(\lambda)$ then by using the empirical Bayes rule indicated above one will suffer a risk which as $n \to \infty$ converges to $R(G)$—the minimum attainable

risk for known G. These rules must therefore be considered "optimal in the limit".

7. Areas of applications of the methods discussed in Sections 5 and 6. The following are some examples where the above described rules seem profitable. Many others could be given.

A. *Medical survey.* It is quite common practice to assume that the seriousness of a disease can be expressed in terms of a parameter λ, such that the higher the value of λ the more serious the disease. The parameter λ for an individual cannot be measured directly, but the individual is subjected to r independent tests the result of each being either "positive" or "negative", and λ stands for the probability of a positive result. In this example it is not unrealistic to assume that λ in the population as a whole (either the entire population or just the population connected with a particular medical center) is a random variable with some unknown distribution G. The model described above has been considered by Neyman [9] and Chiang [2]. In the above model it seems reasonable to assume that there exists a value λ^* such that an individual having a value of λ greater than λ^* must be classified as sick. Loss function (24) seems especially adequate for the present example, and since it seems a more serious mistake to classify a sick person as healthy than to classify a healthy person as sick, a rather low value of λ^* should be chosen. In order to classify each of the patients as sick or healthy, one could here use the empirical Bayes rule given in (36), where x_i denotes the number of positive outcomes on the r tests of the ith individual.

B. *Quality control.* Consider lots containing N items. In order to decide whether a lot should be accepted or not it is customary to sample r items from it, and to accept it when the number of defectives in the sample does not exceed some specified constant c. Let λ be the proportion of defectives in the lot. It has become customary to consider λ as a *random variable* which varies from lot to lot, and is distributed according to distribution function G. Considering the costs of sampling and of rejection and acceptance as functions of λ it is possible to determine, for given N and G, optimal sample size r and constant c. (This problem was raised by Barnard in [1], and solved by Guthrie and Johns in [3] and by Hald in [4]. See also Wetherill [17].) However, since G will usually be unknown this theoretical solution has only limited practical value. We shall instead show, that for fixed sample size r, an empirical Bayes rule can be found. Let A_0 denote the action of accepting the lot and A_1 the action of rejecting it. A reasonable loss function seems to be

$$L(A_0, \lambda) = a\lambda \qquad a > 0$$
$$L(A_1, \lambda) = b(1 - \lambda) \qquad b > 0$$

(where possibly $a = Na_1$ and $b = Nb_1$), since for each non-defective item which is rejected there is a positive loss, and likewise for each defective item which is accepted there is a positive (possibly different) loss. Since

$$L(A_0, \lambda) - L(A_1, \lambda) = (a + b)\lambda - b \tag{43}$$

it follows that for known λ the best procedure would be to accept the lot if $\lambda < b/(a + b)$, reject it if $\lambda < b/(a + b)$ and take either action when $\lambda = b/(a + b)$. The value $\lambda^* = b/(a + b)$ may thus be called the "break even quality". We shall assume that r is small as compared to the lot size N, and thus assume that the number of defectives in the sample is distributed according to the binomial distribution with parameters r, λ. (This assumption is made also in [4].) Since (43) satisfies (23) a proper empirical Bayes rule is again given in (36) where x_i denotes the number of defectives in the ith sample, and where $\lambda^* = b/(a + b)$. This rule will in the limit be as good as any optimal rule if G were known but the sample was of size $r - 1$ rather than r.

From [3] and [4] it follows that the optimal sample size r (when considering also costs of sampling) is a function of G. Thus a modification of the above rule in which also the sample size r of the nth lot is permitted to depend upon previous available information, seems desirable.

C. *College entrance examinations.* Suppose students arrive sequentially and are submitted to a college entrance test, and decision about their admittance is made according to their achievement on the test. It is reasonable to assume that each student has an ability value θ which cannot be measured directly, but that a student's test score is a normal variable with mean θ and some standard deviation σ which is considered fixed for all students.

A student should be admitted if and only if his θ value exceeds some fixed value θ^*. Suppose the loss (e.g. to the Nation) for not admitting a student with $\theta > \theta^*$ is $c[\exp(\theta/\sigma^2) - \exp(\theta^*/\sigma^2)]$ and the loss (e.g. to the college) for admitting a student with $\theta \leqq \theta^*$ is $c[\exp(\theta^*/\sigma^2) - \exp(\theta/\sigma^2)]$, and there is no loss for a correct decision. The parameter θ can reasonably be considered as a random variable, with some (unknown) apriori distribution among college applicants. Thus from Example 2 of Section 6 it follows that a reasonable procedure would be to admit the nth student if and only if his score x_n and the previous scores $\mathbf{x}_{n-1}$ are such that

$$e^{x_n/\sigma^2} \frac{F_{n-1}(x_n + 1 + h_{n-1}) - F_{n-1}(x_n + 1 - h_{n-1})}{F_{n-1}(x_n + h_{n-1}) - F_{n-1}(x_n - h_{n-1})} > e^{(\theta^* - \frac{1}{2})/\sigma^2}$$

where h_n and F_n are given in (39) and (40).

One may of course raise objections to letting the decision about admittance to college depend upon the scores of the students applying earlier, though this seems to be common practice.

In the case where students do not arrive sequentially (i.e. the decisions about admittance are made only after all students have been examined) a modification of the above rule is advantageous. This modification is to treat each of the n students as if he was the last one to apply, and thus for each student one bases the estimate (39) of f_G on all $n - 1$ other students. This approach is better in the sense that with the above modification the average risk over the n individual decisions converges more rapidly to $R(G)$ than the corresponding average risk for the sequential rule. A simplification of the above rule, with the same asymptotic properties and which requires much less computations is to let the estimate

of f_G for each decision be based on *all* n observations, i.e. the ith student is admitted if and only if x_i is such that

$$e^{x_i/\sigma^2}\frac{F_n(x_i+1+h_n)-F_n(x_i+1-h_n)}{F_n(x_i+h_n)-F_n(x_i-h_n)} > e^{(\theta^*-\frac{1}{2})/\sigma^2}.$$

Though the above modifications were discussed only in connection with the particular example C, it is obvious that similar modifications for any of the rules discussed in Sections 5 and 6 are called for whenever one is in the non-sequential rather than in the sequential situation, i.e. whenever the individual decisions have to be made only after all n observations are at hand. The asymptotic optimality properties remain unchanged.

8. Relation to the compound decision problem. The compound decision problem deals with the situation in which one is confronted with n individual decisions about some unknown parameters $\lambda_1, \lambda_2, \cdots, \lambda_n = \mathfrak{L}_n$. In this case, however, $\mathfrak{L}_n$ is considered as an unknown sequence of constants, and not as a realization of n independent random variables Λ_i with distribution G. Thus the Bayes risk is of no interest of its own. Suppose however that before the n decisions have to be made the statistician knows the values of the parameters λ_i, $i = 1, \cdots, n$, but not their order, i.e. the function

$$G_n(\lambda) = (1/n)\ (\text{number of indices}\quad i, i = 1, \cdots, n \quad \text{for which } \lambda_i \leqq \lambda)$$

is known. Then, if the statistician at each decision uses the rule φ_{G_n} which is Bayes with respect to G_n, his *average* risk on the n individual decisions will be $R(G_n)$. It is thus natural to ask, whether also when G_n is *unknown*, one can devise a decision rule $T_n = (t_1, t_2, \cdots, t_n)$ where $0 \leqq t_i \leqq 1$ (and $t_i = t_i(\mathbf{x}_i)$ or $t_i = t_i(\mathbf{x}_n)$ according to whether one is in the sequential or in the non-sequential situation), such that if t_i is used to decide on λ_i one has

$$[R(T_n, \mathfrak{L}_n) - R(G_n)] \to 0 \tag{44}$$

where $R(T_n, \mathfrak{L}_n) = \sum_{i=1}^n R(t_i, \lambda_i)/n$, (see (1)) whatever be the infinite sequence $\mathfrak{L} = \lambda_1, \lambda_2, \cdots$.

When the decision problem is one of testing a simple hypothesis versus a simple alternative rules have been devised for which (44) holds, both for the nonsequential (see [5]) as well as for the sequential (see [15] and [16]) situation. These rules constitute also optimal empirical Bayes rules, and are indicated in Section 3. This is intuitively obvious, since in the Bayesian situation $G_n \to G$ with probability one, and one would hope that also $R(G_n) \to R(G)$.

Usually it will be easier to find empirical Bayes solutions than to find compound decision rules satisfying (44), and the possibility that empirical Bayes rules constitute also "optimal" compound decision rules was stated by Robbins in [11] p. 147. To the best of the present author's knowledge no "optimal" compound decision rules have been exhibited except for the simple versus simple hypothesis testing problem. The author believes that the rules exhibited in

Section 5 and 6 of the present paper constitute "optimal" compound decision rules for the composite hypothesis testing problems considered, at least for the nonsequential compound case (when these rules are modified according to the last paragraph of Section 7), when only mild restrictions are put on the sequence λ. This problem still awaits a rigorous investigation.

Acknowledgment. This research was initiated while the author was at Columbia University and benefited from many stimulating discussions with Professor Herbert Robbins, which she gratefully acknowledges. Application of the methods to medical survey was suggested by Professor J. Neyman, and the possibility of using the methods in quality control was discussed with Professor A. Hald.

REFERENCES

[1] BARNARD, G. A. (1954). Sampling inspection and statistical decisions. *J. Roy. Statist. Soc. Ser. B* **16** 151–174.

[2] CHIANG, C. L. (1951). On the design of mass medical surveys. *Human Biology* **23** 242–271.

[3] GUTHRIE, D. JR. and JOHNS, M. V. JR. (1959). Bayes acceptance sampling procedures for large lots. *Ann. Math. Statist.* **30** 896–925.

[4] HALD, A. (1960). The compound hypergeometric distribution and a system of single sampling inspection plans based on prior distributions and costs. *Technometrics* **2** 257–340.

[5] 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.

[6] JOHNS, M. V. JR. (1956). Contributions to the theory of empirical Bayes procedures in statistics. Doctoral thesis at Columbia University.

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

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

[9] NEYMAN, J. (1947). Outline of statistical treatment of the problem of diagnosis. *Public Health Reports* **62** 1449–1456.

[10] PARZEN, E. (1962). On estimation of a probability density function and mode. *Ann. Math. Statist.* **33** 1065–1076.

[11] ROBBINS, H. (1951). Asymptotic subminimax solutions of compound statistical decision problems. *Proc. Second Berkeley Symp. Math. Statist. and Prob.* Univ. of California Press 131–148.

[12] ROBBINS, H. (1955). An empirical Bayes approach to statistics. *Proc. Third Berkeley Symp. Math. Statist. and Prob.* Univ. of California Press 157–163.

[13] ROBBINS, H. (1963). The empirical Bayes approach to testing statistical hypotheses. To appear in *Rev. Inst. Internat. Statist.* No. 2.

[14] ROSENBLATT, M. (1956). Remarks on some non-parametric estimates of a density function. *Ann. Math. Statist.* **27** 832–837.

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

[16] SAMUEL, E. (1962). Strong convergence of the losses of certain decision rules for the sequential compound decision problem. To appear.

[17] WETHERILL, G. B. (1960). Some remarks on the Bayesian solution of the single sample inspection scheme. *Technometrics* **2** 341–352.