

EMPIRICAL BAYES APPROACH

by 1264

KO-CHUN WANG

B. A., National Taiwan University, 1962

---

A MASTER'S REPORT

submitted in partial fulfillment of the  
requirements for the degree

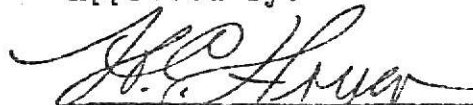
MASTER OF SCIENCE

Department of Statistics and Computer Science

KANSAS STATE UNIVERSITY  
Manhattan, Kansas

1970

Approved by:

  
Major Professor

LD  
2668  
R4  
1970  
W35

TABLE OF CONTENTS

	Page
1. INTRODUCTION . . . . .	1
2. EMPIRICAL BAYES APPROACH TO STATISTIC ESTIMATION . .	2
2.1 Bayes Estimator with Apriori Distribution . .	2
2.2 Empirical Distribution Function . . . . .	4
2.3 Empirical Bayes Approach . . . . .	6
3. EMPIRICAL BAYES APPROACH TO STATISTICAL DECISION PROBLEMS . . . . .	11
3.1 Notation . . . . .	12
3.2 Bayes Decision Function: Known $G$ . . . . .	12
3.3 Empirical Bayes Decision Procedure . . . . .	14
3.4 Application--Poisson Case . . . . .	19
3.5 Remarks . . . . .	22
4. EMPIRICAL BAYES APPROACH TO TESTING HYPOTHESIS . . .	22
4.1 Testing Simple Hypotheses: Known $p$ . . . . .	22
4.2 Testing Simple Hypotheses: Unknown $p$ . . . . .	25
4.3 Testing Composite Hypotheses: Known $G$ . . . . .	30
4.4 Testing Composite Hypotheses: Unknown $G$ . . . . .	32
4.5 Application--Normal Case . . . . .	34
4.6 Remarks . . . . .	38
5. EXTENSION: SEQUENTIAL HYPOTHESIS TESTS FOR STATIONARY PROCESSES . . . . .	39
5.1 Definitions and Notation . . . . .	39
5.2 Bayes Solution: Known $p$ . . . . .	41
5.3 Empirical Bayes Solution: Unknown $p$ . . . . .	43
6. CONCLUSION . . . . .	45
ACKNOWLEDGMENT . . . . .	48
REFERENCES . . . . .	49

## 1. INTRODUCTION

Consider a sequence of observations generated by a statistical experiment which depends on an unknown parameter, and, in addition, consider as given that the parameter is a random variable but with an unknown distribution. It will usually be advantageous to treat the whole sequence of observations as a single entity rather than to treat each component separately. It was with this idea in mind that Robbins (1955) first discussed the use of apriori observations to approximate Bayes procedures, and hence to establish the "asymptotically optimal" statistical solutions. In that paper Robbins described such approximations as "empirical Bayes procedures".

It has been proved that whenever a statistical experiment comes to us with such a sequence of observations, the empirical Bayes approach offers certain advantages over any other approach which either ignores the fact that the unknown parameter is itself a random variable, or assumes a personal or subjective probability distribution of the parameter not subject to change with experience.

Since the time of Robbins' initial work--which Neyman (1962) admired as the first breakthrough in statistical theory during the past decade--on the utilization of previous experience for statistical inference, the theory of the empirical Bayes approach has been so developed that some of the theories and solutions are ready for practical applications.

This report considers some of the discussions by Robbins (1955, 1963, 1964), Johns (1957, 1961), Neyman (1962), Samuel (1963), and Tainiter (1965). The empirical Bayes approach to statistical estimation is described in Section 2, where for simplicity only discrete random variables are considered. In Section 3, the "asymptotically optimal" empirical Bayes rules and their applications for decision problems are considered. In Section 4, the asymptotically optimal empirical Bayes approach to the testing of hypotheses, together with some examples, is considered. Section 5 is an extension which deals with some recent works by Tainiter. The last section is for concluding remarks.

## 2. EMPIRICAL BAYES APPROACH TO STATISTIC ESTIMATION

### 2.1 Bayes Estimator with Apriori Distribution

Let  $X$  be a random variable with a known probability density function depending on an unknown real parameter  $\Lambda$ , namely,

$$p(x|\lambda) = \Pr[X = x | \Lambda = \lambda].$$

Suppose  $\Lambda$  is itself a random variable with apriori distribution function

$$G(\lambda) = \Pr[\Lambda \leq \lambda],$$

then the marginal p.d.f. of  $X$  is given by



$$P_G(x) = \Pr[X = x] = \int_{\Lambda} p(x|\lambda) dG(\lambda). \quad (2.1)$$

If mean square error is adopted as the measurement of accuracy, then the expected squared deviation of any estimator of  $\Lambda$  of the form  $\Psi(x)$  is

$$\begin{aligned} E[\Psi(x) - \Lambda]^2 &= E\{E[(\Psi(x) - \Lambda)^2 | \Lambda = \lambda]\} \\ &= E\left\{\sum_{\bar{x}} (\Psi(x) - \lambda)^2 p(x|\lambda)\right\} \\ &= \int_{\Lambda} \sum_{\bar{x}} p(x|\lambda) [\Psi(x) - \lambda]^2 dG(\lambda) \\ &= \sum_{\bar{x}} \int_{\Lambda} p(x|\lambda) [\Psi(x) - \lambda]^2 dG(\lambda). \end{aligned} \quad (2.2)$$

This quantity attains its minimum if, for each  $x$  the estimator given by  $\Psi_0 = \Psi_0(x)$  is such that

$$I(x) = \int_{\Lambda} p(x|\lambda) [\Psi_0 - \lambda]^2 dG(\lambda) = \text{minimum.}$$

But for any fixed  $x$  the quantity

$$\begin{aligned} I(x) &= \Psi_0^2 \int_{\Lambda} p dG - 2 \Psi_0 \int_{\Lambda} p \lambda dG + \int_{\Lambda} p \lambda^2 dG \\ &= \left(\Psi_0 - \frac{\int_{\Lambda} p \lambda dG}{\int_{\Lambda} p dG}\right)^2 \int_{\Lambda} p dG + \left[\int_{\Lambda} p \lambda^2 dG - \frac{(\int_{\Lambda} p \lambda dG)^2}{\int_{\Lambda} p dG}\right] \end{aligned} \quad (2.3)$$

is a minimum w.r.t.  $\Psi_0$  when

$$\Psi_0 = \frac{\int_{\Lambda} p \lambda dG}{\int_{\Lambda} p dG}.$$

The minimum value of  $I(x)$  is

$$I_G(x) = \int_{\Lambda} p\lambda^2 dG - \frac{(\int_{\Lambda} p\lambda dG)^2}{\int_{\Lambda} p dG}.$$

Hence the Bayes estimator of  $\Lambda$  corresponding to the apriori distribution function  $G$  of  $\Lambda$  is the random variable  $\Psi_G(x)$  defined by

$$\Psi_G(x) = \frac{\int_{\Lambda} p(x|\lambda)\lambda dG(\lambda)}{\int_{\Lambda} p(x|\lambda) dG(\lambda)}. \quad (2.4)$$

The corresponding minimum expected square error is

$$E [\Psi_G(x) - \Lambda]^2 = \sum_x I_G(x).$$

## 2.2 Empirical Distribution Function

If the apriori distribution function  $G$  is known, then  $\Psi_G$  as defined by (2.4) can be computed. However, even when  $G$  may be assumed to exist, it is usually unknown to the experimenter, and hence  $\Psi_G$  cannot be computed.

Now, suppose that the problem of estimating  $\Lambda$  from an observed value of  $X$  occurs repeatedly and independently. Then let

$$(\lambda_1, x_1), (\lambda_2, x_2), \dots, (\lambda_n, x_n), \dots \quad (2.5)$$

denote a sequence of pairs of random variables, each pair being independent of all the other pairs. The conditional distribution of  $x$ , given  $\lambda$ , is  $p(x|\lambda)$ . Thus the distribution of  $x_n$  depends only on  $\lambda_n$  and for  $\lambda_n = \lambda$  is given by  $p(x|\lambda)$ . The process

$\{\lambda_n\}$  is a sequence of independent, identically distributed random variables having a common apriori distribution  $G$  defined on  $\Lambda$ .

If it is desired to estimate an unknown  $\lambda_n$  from an observed  $x_n$ , and if the previous values  $\lambda_1, \lambda_2, \dots, \lambda_{n-1}$  are known, then we can form the empirical distribution function based on the  $n-1$  independent trials to be

$$G_{n-1}(\lambda) = \frac{\text{number of terms } \lambda_1, \lambda_2, \dots, \lambda_{n-1} \text{ which are } \leq \lambda}{n-1} \quad (2.6)$$

Next, replace the unknown apriori  $G$  in (2.4) by the empirical  $G_{n-1}$  in (2.6) to obtain

$$\psi_n(x) = \frac{\int_{\Lambda} p(x|\lambda) \lambda dG_{n-1}(\lambda)}{\int_{\Lambda} p(x|\lambda) dG_{n-1}(\lambda)}.$$

It has been proved, Glivenko (1933), that as the number of trials increases indefinitely, an empirical distribution uniformly converges to a theoretical distribution with probability 1, i.e.,  $G_{n-1}(\lambda) \rightarrow G(\lambda)$ . Therefore under suitable conditions  $\psi_n$  will tend to  $\psi_G$ , the estimator of  $\Lambda$  defined in (2.4).

In practice, however, it will be unusual for the previous values  $\lambda_1, \lambda_2, \dots, \lambda_{n-1}$  to be known. In most cases these values are unknown and the observations are limited to the values of  $x_1, x_2, \dots, x_n, \dots$ . Thus  $G_{n-1}$  is unknown and  $\psi_n$  is not obtainable. However, it is reasonable to assume that the values  $x_1, x_2, \dots, x_n$  are available when  $\lambda_n$  is to be estimated. Therefore for any fixed  $X$ , the empirical relative

frequency distribution is given by

$$P_n(x) = \frac{\text{number of terms } x_1, x_2, \dots, x_n \text{ which equal } x}{n} \quad (2.7)$$

and tends to  $P_G$  as  $n \rightarrow \infty$  for any apriori distribution function  $G$ . This convergence property is used in developing the empirical Bayes approach which follows.

### 2.3 Empirical Bayes Approach

The object in this approach is to observe the  $x_i$ 's, and, on the basis of these observations, to exhibit an empirical relative frequency function, and hence to estimate the unknown  $G$  and the value of the Bayes estimator of  $\Lambda$ . This, of course, will depend on the nature of the p.d.f.  $p(x|\lambda)$  and on the class of distributions to which the unknown  $G$  is assumed to belong. To the second part of this problem Robbins (1955) proposed the following solution. Find a function of  $x_1, x_2, \dots, x_{n-1}$  and of  $x_n = x$ , say

$$\psi_n(x_1, x_2, \dots, x_{n-1}; x) = \psi_n(x_n) \quad (2.8)$$

such that  $\psi_n(x_n)$  is a consistent estimate of  $\psi_G(x)$ . If more than one such function is available, determine the one which is best in some sense (e.g., the fastest rate of convergence, the best choice for minimizing the deviation from the estimator, etc.). Several examples will be considered to illustrate the preceding ideas.

Consider first the Poisson case. That is, consider

$$p(x|\lambda) = e^{-\lambda} \frac{\lambda^x}{x!}, \quad x = 0, 1, \dots, \quad \lambda > 0.$$

From (2.1), it follows that

$$\begin{aligned} p_G(x) &= \int_{\Lambda} p(x|\lambda) dG(\lambda) \\ &= \left[ \int_0^{\infty} e^{-\lambda} \lambda^x dG(\lambda) \right] / x!. \end{aligned}$$

By (2.4), the Bayes estimator of  $\Lambda$  is

$$\begin{aligned} \psi'_G(x) &= \frac{\int_0^{\infty} e^{-\lambda} \lambda^{x+1} dG(\lambda) / x!}{\int_0^{\infty} e^{-\lambda} \lambda^x dG(\lambda) / x!} \\ &= \frac{(x+1) \int_0^{\infty} e^{-\lambda} \lambda^{x+1} dG(\lambda) / (x+1)!}{\int_0^{\infty} e^{-\lambda} \lambda^x dG(\lambda) / x!} \\ &= (x+1) \frac{p_G(x+1)}{p_G(x)}. \end{aligned}$$

Now, define

$$\psi'_n(x) = (x+1) \frac{p_n(x+1)}{p_n(x)} \quad (2.9)$$

$$= (x+1) \frac{\text{number of terms } x_1, \dots, x_n \text{ which equal } x+1}{\text{number of terms } x_1, \dots, x_n \text{ which equal } x}.$$

Then it can be shown that for any fixed  $x$ ,

$$\lim_{n \rightarrow \infty} \psi_n(x) \longrightarrow \psi_G(x)$$

with probability 1 for any unknown  $G$ . In view of this idea, the computable quantity  $\psi_n(x)$  can be used as an estimate of the unknown  $\lambda_n$  in the sequence (2.5) in the hope that as  $n \rightarrow \infty$

$$E[\psi_n(x) - \lambda_n]^2 \longrightarrow E[\psi_G(x) - \lambda]^2.$$

It is of interest to compare the expected squared deviation of the empirical Bayes estimator  $\psi_n(x)$  with the usual maximum likelihood estimator for the Poisson parameter, say  $X$ , for which

$$E(X - \lambda)^2 = E_{\lambda} E_X[(X - \lambda)^2 | \lambda] = E(\lambda) = \int_0^{\infty} \lambda dG(\lambda).$$

Suppose that  $G$  has gamma distribution function with density

$$G'(\lambda) = \frac{h^b}{\Gamma(b)} \lambda^{b-1} e^{-h\lambda}, \quad \lambda, b, h > 0.$$

Then

$$E(\lambda) = \frac{b}{h}, \quad \text{var}(\lambda) = \frac{b}{h^2}.$$

From (2.4), the Bayes estimator of  $\lambda$  is

$$\begin{aligned} \psi_G(x) &= \frac{\int_{\lambda} p(x|\lambda) \lambda dG(\lambda)}{\int_{\lambda} p(x|\lambda) dG(\lambda)} \\ &= \frac{\int_{\lambda} (e^{-\lambda} \lambda^{x+1}) (c \lambda^{b-1} e^{-h\lambda}) d\lambda / x!}{\int_{\lambda} (e^{-\lambda} \lambda^x) (c \lambda^{b-1} e^{-h\lambda}) d\lambda / x!} \end{aligned}$$

$$= \frac{\int_{\lambda} e^{-\lambda(1+h)} \lambda^{x+b+1-1} d\lambda}{\int_{\lambda} e^{-\lambda(1+h)} \lambda^{x+b-1} d\lambda},$$

where  $c = h^h / \Gamma(b)$ . Writing both denominator and numerator as integrals of gamma densities gives

$$\begin{aligned} \psi_G(x) &= \frac{\frac{\Gamma(x+b+1)}{(1+h)^{x+b+1}} \int_{\lambda} \frac{(1+h)^{x+b+1}}{(x+b+1)} e^{-\lambda(1+h)} \lambda^{x+b+1-1} d\lambda}{\frac{\Gamma(x+b)}{(1+h)^{x+b}} \int_{\lambda} \frac{(1+h)^{x+b}}{\Gamma(x+b)} e^{-\lambda(1+h)} \lambda^{x+b-1} d\lambda} \\ &= \frac{x+b}{1+h}. \end{aligned}$$

The expected squared deviation of the empirical Bayes estimator is then

$$\begin{aligned} E \left[ \psi_G(x) - \lambda \right]^2 &= E \left[ \left( \frac{x+b}{1+h} \right) - \lambda \right]^2 \\ &= E_{\lambda} \left[ E \left\{ \left( \frac{x+b}{1+h} \right)^2 - 2\lambda \left( \frac{x+b}{1+h} \right) + \lambda^2 \right\} | \lambda \right] \\ &= E_{\lambda} \left[ \frac{1}{(1+h)^2} [\lambda + \lambda^2 + 2b\lambda + b^2] - \frac{2\lambda}{1+h} (\lambda + b) + \lambda^2 \right] \\ &= \frac{b}{h(1+h)}. \end{aligned}$$

Thus the relative efficiency is

$$\frac{E[\psi_G(x) - \lambda]^2}{E[x - \lambda]^2} = \frac{1}{1 + h}.$$

As a second example consider the geometric distribution

$$p(x|\lambda) = (1 - \lambda)\lambda^x, \quad x = 0, 1, \dots, \quad 0 \leq \lambda < 1.$$

Then (2.1) gives

$$p_G(x) = \int_0^1 (1 - \lambda)\lambda^x dG(\lambda)$$

and (2.4) yields

$$\begin{aligned} \psi_G(x) &= \frac{\int_0^1 (1 - \lambda)\lambda^{x+1} dG(\lambda)}{\int_0^1 (1 - \lambda)\lambda^x dG(\lambda)} \\ &= \frac{p_G(x+1)}{p_G(x)}. \end{aligned}$$

This suggests that the estimator be defined as

$$\psi_n(x) = \frac{\text{number of terms } x_1, \dots, x_n \text{ which are equal to } x+1}{\text{number of terms } x_1, \dots, x_n \text{ which are equal to } x}. \quad (2.10)$$

Note that as the number of earlier observations  $x_1, x_2, \dots, x_{n-1}$  increases, the estimators such as (2.9) and (2.10) will be almost as accurate as the estimate  $\psi_G$  which requires complete knowledge of the apriori distribution  $G$ . Further, only one empirical Bayes estimator of  $\lambda_n$  has been studied for each of the two examples discussed above while more than one such estimator



may exist for each distribution. The gain in precision produced by the empirical Bayes approach over the estimate from any other procedure depends upon the nature of  $G$  and can be considerable.

Most of the ideas discussed in this section are for discrete random variables; however, almost all the results can be generalized to the continuous cases. Also, the discussion has been concerned only with the second part of the problem, i.e., obtaining an approximate value of an empirical Bayes estimate of  $\lambda_n$ . The first part of the problem, obtaining an approximation to the unknown  $G$ , still awaits a satisfactory solution. To pure Bayesians  $G$  will always be known by introspection, and to pure non-Bayesians  $G$  will not even exist. Thus to these individuals this question does not exist. Taking the position that  $G$  exists but always remains unknown, Robbins (1964) presented a possible way—a special case of the "minimum distance" method of J. Wolfowitz—for constructing an empirical approximation to the unknown  $G$ .

### 3. EMPIRICAL BAYES APPROACH TO STATISTICAL DECISION PROBLEMS

The empirical Bayes approach is applicable to decision problems for which repeated and independent observations of a random variable, whose distribution depends upon a parameter with a fixed but unknown apriori distribution, are available. Not all decision problems come to us with such a sequence, but, when they do, the empirical Bayes approach offers certain

advantages over any other approach which ignores the existence of the apriori distribution subject to change with experience.

### 3.1 Notation

Consider an observable random variable  $x$  defined in a space  $X$  on which a  $\sigma$ -finite measure  $\mu$  is defined. Given  $\lambda$  (real parameter),  $x$  has conditional p.d.f.  $f_\lambda$  with respect to  $\mu$ . The parameter  $\lambda$  is defined in  $\Lambda$ . Further,  $\lambda$  has an apriori distribution  $G$  which may or may not be known to us. Let  $A$  denote an action space with generic element  $a$ . Let  $L[a, \lambda] \geq 0$  represent the loss in taking action  $a$  when the parameter is  $\lambda$ .

### 3.2 Bayes Decision Function: Known G

Consider a decision function  $t$  defined on  $X$  and with values in  $A$ , such that for any  $x$  the corresponding action  $t(x)$  will be taken, and hence incur the loss  $L[t(x), \lambda]$ . For any  $t$  it follows that the expected loss is given by

$$R(t, \lambda) = \int_X L[t(x), \lambda] f_\lambda(x) d\mu(x) .$$

Further, the overall expected loss (i.e., Bayes risk relative to  $G$ ) is

$$R(t, G) = \int_\Lambda R(t, \lambda) dG(\lambda) . \quad (3.1)$$

Let

$$\phi_G(a, x) = \int_\Lambda L[a, \lambda] f_\lambda(x) dG(\lambda), \quad (3.2)$$

then (3.1) can be written

$$\begin{aligned}
 R(t, G) &= \int_{\Lambda} \int_{\mathbf{X}} L[t(x), \lambda] f_{\lambda}(x) d_{\mu}(x) dG(\lambda) \\
 &= \int_{\mathbf{X}} \int_{\Lambda} L[t(x), \lambda] f_{\lambda}(x) dG(\lambda) d_{\mu}(x) \\
 &= \int_{\mathbf{X}} \phi_G(t(x), x) d_{\mu}(x) .
 \end{aligned}$$

Assume that there exists a decision function  $t_G$  such that for every  $x$ ,

$$\phi_G(t_G(x), x) = \min_a \phi_G(a, x) , \quad (3.3)$$

then for any decision function  $t$ ,

$$R(t_G, G) = \int_{\mathbf{X}} \min_a \phi_G(a, x) d_{\mu}(x) \leq R(t, G) .$$

Defining

$$R(G) = R(t_G, G) = \int_{\mathbf{X}} \phi_G(t_G(x), x) d_{\mu}(x)$$

implies that

$$R(G) = \min_t R(t, G) . \quad (3.4)$$

Thus decision function  $t_G$  satisfying (3.3) minimizes the Bayes risk (3.1) and is called the Bayes decision function relative to  $G$ . If  $G$  is known,  $t_G$  is available and the minimum possible Bayes risk  $R(G)$  is attainable.

### 3.3 Empirical Bayes Decision Procedure

Since  $G$  is assumed to exist,  $R(t, G)$  is an appropriate criterion for measuring the performance of any decision function  $t$ . Also, since  $G$  is assumed to be unknown,  $t_G$  is not directly available. Therefore a method is needed for determining  $R(G)$  in (3.4) when  $G$  is unknown. To find a solution, consider an experiment which deals with a sequence of pairs of random variables as in (2.5), i.e.,

$$(\lambda_1, x_1), (\lambda_2, x_2), \dots, (\lambda_n, x_n) \dots$$

The conditional p.d.f. of  $x_n$  given that  $\lambda_n = \lambda$  is  $f_\lambda$ . The observations are limited to the values of  $x_1, x_2, \dots, x_n, \dots$ . It is assumed that at the time when the decision about  $\lambda_n$  is to be made, the values of  $x_1, x_2, \dots, x_n$  are known. This suggests that the decision about  $\lambda_n$  is to be based on a function of  $x_1, x_2, \dots, x_{n-1}$  and  $x_n = x$  of the form

$$t_n(x) = t_n(x_1, x_2, \dots, x_{n-1}; x). \quad (3.5)$$

Taking action  $t_n(x_n) \in A$  will incur the loss  $L[t_n(x_n), \lambda_n]$ . Since the successive terms in the sequence are independent and with the same structure, it seems reasonable to use a fixed decision function  $t$  for each  $n$ . The reason for using  $t_n$  instead of a fixed  $t$  is that as  $n$  increases, the sample points  $x_1, x_2, \dots, x_n$  will contain more information about the unknown  $G$ . Thus for large  $n$ ,  $t_n$  will be close to the optimal but unknown  $t_G$ .

Suppose an "empirical decision procedure" is defined to be

a sequence  $T = \{t_n\}$  of the form (3.5) with values defined in action space  $A$ . Then for a given  $T$ , the expected loss associated with  $t_n \in T$  for the decision about  $\lambda_n$  given  $x_1, x_2, \dots, x_n$  is

$$\begin{aligned} R_n(t, G) &= \int_{\mathbf{x}} \int_{\Lambda} L[t_n(\mathbf{x}), \lambda] f_{\lambda}(\mathbf{x}) dG(\lambda) d_{\mu}(\mathbf{x}) \\ &= \int_{\mathbf{x}} \phi_G(t_n(\mathbf{x}), \mathbf{x}) d_{\mu}(\mathbf{x}) . \end{aligned}$$

The overall expected loss for  $T$  is

$$R_n(T, G) = \int_{\mathbf{x}} E \phi_G(t_n(\mathbf{x}), \mathbf{x}) d_{\mu}(\mathbf{x}) \quad (3.6)$$

where the symbol  $\mathbf{x}$  plays the role of a dummy variable of integration and not a random variable, and  $E$  denotes expectation with respect to the  $n-1$  independent apriori random variables  $x_1, x_2, \dots, x_{n-1}$  which have the common p.d.f. with respect to  $\mu$  on  $X$  given by

$$f_G(\mathbf{x}) = \int_{\Lambda} f_{\lambda}(\mathbf{x}) dG(\lambda) . \quad (3.7)$$

Now, from (3.4) and (3.6) it is obvious that  $R_n(T, G) \geq R(G)$ .

Before proceeding with further discussion about  $R_n(T, G)$ , we now quote a definition credited to Robbins (1964).

DEFINITION:  $T$  is said to be asymptotically optimal (a.o.) relative to  $G$  if

$$\lim_{n \rightarrow \infty} R_n(T, G) = R(G) . \quad (3.8)$$

Now the problem is to find the empirical decision sequence  $T$ , which for large  $n$  is in some sense "best" relative to the

unknown  $G$ . To answer a part of this question, it is necessary to take a further look at some generalities on asymptotically optimal. Rewriting (3.8) we have, for  $T$  to be a.o. relative to  $G$ , the following condition

$$\lim_{n \rightarrow \infty} \int_X E\phi_G(t_n(x), x) d\mu(x) = \int_X \phi_G(t_G(x), x) d\mu(x), \quad (3.9)$$

where  $\{E\phi_G(t_n(x), x)\}$  is a sequence of measuring functions and  $x \in X$  is a set of random variables on which a  $\sigma$ -finite measure  $\mu$  is defined. In order to prove  $T = \{t_n\}$  is a.o. relative to  $G$ , by Lebesgue's theorem on dominated convergence it suffices to prove the following equations

$$(a) \quad \lim_{n \rightarrow \infty} E\phi_G(t_n(x), x) = \phi_G(t_G(x), x)$$

$$(b) \quad E\phi_G(t_n(x), x) \leq H(x) \quad (\text{all } n)$$

where

$$\int_X H(x) d\mu(x) < \infty. \quad (3.10)$$

The main problem is (a); we shall summarily dispose of (b) by assuming

$$(c) \quad \int_{\Lambda} L(\lambda) dG(\lambda) < \infty,$$

where we have set

$$0 \leq L(\lambda) = \sup_a L[a, \lambda] < \infty.$$

By letting

$$H(x) = \int_{\Lambda} L(\lambda) f_{\lambda}(x) dG(\lambda) \geq 0,$$

we have by (3.2) for any  $T$ ,

$$\phi_G(t_n(x), x) \leq H(x) \quad (\text{all } n) \quad (3.11)$$

and from (3.10),

$$\begin{aligned} \int_x H(x) d\mu(x) &= \int_{\Lambda} L(\lambda) \int_x f_{\lambda}(x) d\mu(x) dG(\lambda) \\ &= \int_{\Lambda} L(\lambda) dG(\lambda) < \infty, \end{aligned} \quad (3.12)$$

and (3.11) and (3.12) imply that (b) holds. Moreover, from (3.12), it follows that  $H(x) < \infty$ , and hence to prove that (a) holds it will suffice to prove that

$$(d) \quad P \lim_{n \rightarrow \infty} \phi_G(t_n(x), x) = \phi_G(t_G(x), x)$$

where  $P \lim$  means limit in probability. Therefore, (c) and (d) insure that  $T$  is a.o. relative to  $G$ .

Let  $a_0$  be an arbitrary fixed element of  $A$  and define

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

and

$$L_0(x) = \int_{\Lambda} L[a_0, \lambda] f_{\lambda}(x) dG(\lambda),$$

so that under (c) we have

$$\begin{aligned} \Delta_G(a, x) &= \int_{\Lambda} L[a, \lambda] f_{\lambda}(x) dG(\lambda) - \int_{\Lambda} L[a_0, \lambda] f_{\lambda}(x) dG(\lambda) \\ &= \phi_G(a, x) - L_0(x) \end{aligned}$$

which is equivalent to

$$\phi_G(a, x) = L_0(x) + \Delta_G(a, x) . \quad (3.13)$$

Suppose we can find a sequence of functions

$$\Delta_n(a, x) = \Delta_n(x_1, x_2, \dots, x_{n-1}; a, x) \quad (3.14)$$

such that

$$P \lim_{n \rightarrow \infty} \sup_a |\Delta_n(a, x) - \Delta_G(a, x)| = 0. \quad (3.15)$$

Let

$$t_n(x) = t_n(x_1, \dots, x_{n-1}; x) = \text{any element } \bar{a} \in A \quad (3.16)$$

such that for  $\epsilon_n > 0$ ,

$$\Delta_n(\bar{a}, x) \leq \inf_a \Delta_n(a, x) + \epsilon_n .$$

Then by (3.3) and (3.13),

$$\begin{aligned} 0 &\leq \Delta_G(t_n(x), x) - \Delta_G(t_G(x), x) & (3.17) \\ &= [\Delta_G(t_n(x), x) - \Delta_n(t_n(x), x)] \\ &\quad + [\Delta_n(t_n(x), x) - \Delta_n(t_G(x), x)] \\ &\quad + [\Delta_n(t_G(x), x) - \Delta_G(t_G(x), x)] . \end{aligned}$$

From (3.15) we see that given any  $\epsilon > 0$ ,  $n$  can be chosen such that the right-hand side of (3.17) will be less than or equal to  $\epsilon + \epsilon_n + \epsilon$ ; thus

$$P \lim_{n \rightarrow \infty} \Delta_G(t_n(x), x) = \Delta_G(t_G(x), x)$$

which by (3.13) implies (d). The foregoing can be summed up in



the following theorem.

THEOREM 3.1: Let  $G$  be such that  $\int_{\Lambda} L(\lambda) dG(\lambda) < \infty$  holds, let  $\Delta_n(a, x)$  be a sequence of functions of the form (3.14) and such that (3.15) holds, and define  $T = \{t_n\}$  by (3.16). Then  $T$  is a.o. relative to  $G$ .

When the action space  $A$  is finite, especially when  $A$  is a set of two elements, the following corollary, due to Robbins (1964), provides a solution to the hypothesis testing problem.

COROLLARY 3.1: Let  $A = \{a_0, a_1\}$ , let  $G$  be such that

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

and let  $\Delta_n(x) = \Delta_n(x_1, \dots, x_{n-1}; x)$  be such that

$$P \lim_{n \rightarrow \infty} \Delta_n(x) = \Delta_G(x) = \int_{\Lambda} [L(a_1, \lambda) - L(a_0, \lambda)] f_{\lambda}(x) dG(\lambda). \quad (3.18)$$

Define

$$\begin{aligned} t_n(x) &= a_0, \text{ if } \Delta_n(x) \geq 0, \\ &= a_1, \quad \Delta_n(x) < 0. \end{aligned}$$

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

### 3.4 Application--Poisson Case

Consider the problem of testing a one-side null hypothesis  $H_0: \lambda \leq \lambda^*$  against the alternative hypothesis  $H_1: \lambda > \lambda^*$  for the value of a Poisson parameter  $\lambda$ . Then  $A = \{a_i, i = 0, 1\}$  where  $a_i$  denotes the action "accept  $H_i$ " and

$$f_{\lambda}(x) = \frac{e^{-\lambda} \lambda^x}{x!}, \quad (\lambda > 0; X = 0, 1, 2, \dots),$$

where a  $\sigma$ -finite measure  $\mu$  is defined on  $X$ . Let the loss functions be

$$L[a_0, \lambda] = \text{loss in taking action } a_0 = \begin{cases} 0 & \text{if } \lambda \leq \lambda^*, \\ \lambda - \lambda^* & \text{if } \lambda > \lambda^*, \end{cases}$$

$$L[a_1, \lambda] = \text{loss in taking action } a_1 = \begin{cases} \lambda^* - \lambda & \text{if } \lambda \leq \lambda^*, \\ 0 & \text{if } \lambda > \lambda^*. \end{cases}$$

Thus

$$L[a_1, \lambda] - L[a_0, \lambda] = \lambda^* - \lambda$$

and from (3.18),

$$\begin{aligned} \Delta_G(x) &= \int_{\lambda} [L[a_1, \lambda] - L[a_0, \lambda]] f_{\lambda}(x) dG(\lambda) \\ &= \int_0^{\infty} (\lambda^* - \lambda) \frac{e^{-\lambda} \lambda^x}{x!} dG(\lambda) \\ &= \lambda^* \int_0^{\infty} \frac{e^{-\lambda} \lambda^x}{x!} dG(\lambda) - (x+1) \int_0^{\infty} \frac{e^{-\lambda} \lambda^{x+1}}{(x+1)!} dG(\lambda). \end{aligned}$$

Now by (3.7),

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

so that we can write

$$\Delta_G(x) = \lambda^* f_G(x) = (x+1) f_G(x+1). \quad (3.19)$$

Define

$$\begin{aligned}\delta(x, y) &= 1 && \text{if } x = y, \\ &= 0 && \text{if } x \neq y,\end{aligned}$$

and consider the expression

$$U_n(x) = U_n(x_1, \dots, x_n; x) = n^{-1} \sum_{j=1}^n \delta(x, x_j).$$

Noting that

$$E \delta(x, x_j) = P[x_j = x] = f_G(x).$$

By the strong law of large numbers,

$$P \lim_{n \rightarrow \infty} U_n(x) = f_G(x) \quad (x = 0, 1, 2, \dots). \quad (3.20)$$

(3.19) and (3.20) together suggest the following equation.

$$\Delta_n(x) = \lambda^* U_n(x) - (x+1)U_n(x+1). \quad (3.21)$$

It follows that for  $x = 0, 1, 2, \dots$ ,

$$P \lim_{n \rightarrow \infty} \Delta_n(x) = \lambda^* f_G(x) - (x+1)f_G(x+1) = \Delta_G(x).$$

Set

$$\begin{aligned}t_n(x) &= a_0 && \text{if } \lambda^* U_n(x) - (x+1)U_n(x+1) \geq 0, \\ &= a_1 && \text{otherwise.}\end{aligned}$$

Then by the Corollary 3.1,  $T$  is a.o. relative to every  $G$  such that

$$\int_0^\infty \lambda dG(\lambda) < \infty.$$

### 3.5 Remarks

The relation (3.19) was basic to the construction of (3.21) of a sequence  $\Delta_n(x)$  satisfying (3.18). This special case for the Poisson distribution and the loss structure specified illustrates the application of Corollary 3.1. Johns (1961), Robbins (1963), and Samuel (1963) discuss the application of this Corollary to more general loss models and to many of the most common discrete and continuous parametric distributions of statistics.

When an a.o.  $T$  does not exist as may be the case, or when it does exist but  $R_n(T, G)$  converges to  $R(G)$  too slowly, the empirical Bayes approach is still useful under certain circumstances by using a  $T$  which is "asymptotically subminimax". This procedure was first established by Robbins (1964) and developed by R. Cogburn, Samuel and Robbins.

## 4. EMPIRICAL BAYES APPROACH TO TESTING HYPOTHESES

### 4.1 Testing Simple Hypotheses: Known $p$

Let a parameter  $\lambda$ , representing the unknown "state of Nature" in some statistical experiment, take only two values labeled "0" and "1". Now suppose an observable sequence of random variables  $x \in X$  has a specified probability distribution  $P_\lambda$ , i.e.,

$$P_{\lambda} = \begin{cases} P_0, & \text{when } \lambda = 0, \\ P_1, & \text{when } \lambda = 1. \end{cases} \quad (4.1)$$

For every single observation of  $x$ , it is required to take one of the two actions  $a_0$  and  $a_1$ ,  $a_0$  being the more appropriate when  $\lambda = 0$ , and  $a_1$  the more appropriate when  $\lambda = 1$ . Again let  $L[i, \lambda]$  denote the loss incurred when action  $a_i$  is taken; suppose that

$$L[0, 0] = 0, \quad L[0, 1] = c_1,$$

$$L[1, 0] = c_0, \quad L[1, 1] = 0,$$

where  $c_0$  and  $c_1$  are given positive constants.

The choice of action depends upon a decision function  $t = t(x)$  with values 0 and 1 defined on the sample space  $X$ . Let  $L[t(x), \lambda]$  represent the loss function associated with every decision; it can be written as

$$\begin{aligned} L[t(x), \lambda] &= L[0, \lambda] - t(x)\{L[0, \lambda] - L[1, \lambda]\} \\ &= L[0, \lambda] - t(x)b(\lambda), \end{aligned} \quad (4.2)$$

where

$$b(\lambda) = L[0, \lambda] - L[1, \lambda].$$

Suppose  $\lambda$  is itself a random variable with apriori distribution function

$$\begin{aligned} P[\lambda = 1] &= p, \\ P[\lambda = 0] &= 1 - p, \end{aligned} \quad 0 \leq p \leq 1. \quad (4.3)$$

For any  $t$  it follows that the overall expected loss (i.e., Bayes risk relative to  $G$ ) is

$$R(t, p) = E\{L[t(x), \lambda]\}, \quad (4.4)$$

where  $E$  denotes expectation w.r.t. the joint distribution of the pair of random variables  $(\lambda, x)$ . Using (4.2) we can write

$$\begin{aligned} R(t, p) &= E\{L[t(x), \lambda]\} \\ &= E\{L[0, \lambda]\} - E\{t(x)L[0, \lambda] - t(x)L[1, \lambda]\} \\ &= pc_1 - pc_1 \int_X t(x) dP_1(x) + (1-p)c_0 \int_X t(x) P_0(x). \end{aligned} \quad (4.5)$$

The family of distributions  $P_\lambda$  of  $x$  can be regarded in terms of their p.d.f.  $f_\lambda$  w.r.t. some measure  $\mu$  on  $X$ , so that (4.5) can be written

$$\begin{aligned} R(t, p) &= pc_1 - pc_1 \int_X t(x) f_1(x) d_\mu(x) + (1-p)c_0 \int_X t(x) f_0(x) d_\mu(x) \\ &= pc_1 - \int_X t(x) \psi_p(x) d_\mu(x), \end{aligned} \quad (4.6)$$

where

$$\psi_p(x) = pc_1 f_1(x) - (1-p)c_0 f_0(x). \quad (4.7)$$

Since  $t(x)$  takes only the values 0 and 1, it is clear that for any given  $p$ ,  $R(t, p)$  is minimized by using the decision function  $t_p$  defined by

$$t_p(x) = 1, \quad \text{if } \psi_p(x) \geq 0, \quad (4.8)$$

= 0 otherwise.

The term  $t_p(x)$  is called a Bayes decision function corresponding to a given  $p$ . The minimum value of  $R(t, p)$  being

$$R(p) = \min_t R(t, p) = R(t_p, p) = c_1 p - \int_x [\psi_p(x)]^+ d_\mu(x), \quad (4.9)$$

where

$$\begin{aligned} [\psi_p(x)]^+ &= [\psi_p(x)] \quad \text{if } [\psi_p(x)] \geq 0, \\ &= 0 \quad \text{if } [\psi_p(x)] < 0. \end{aligned}$$

$R(p)$  is called the Bayes envelope function. When  $p$  is known we can use  $t_p$  and thereby incur the minimum possible Bayes risk  $R(p)$ .

#### 4.2 Testing Simple Hypothesis: Unknown $p$

If the apriori distribution function  $p$  is unknown, as is usually the case, then  $\psi_p$  is not a computable function and  $t_p$  is not directly available. Suppose, however, that the same decision problem occurs repeatedly and independently. Then a sequence of independent pairs of random variables can be written by

$$(\lambda_1, x_1), (\lambda_2, x_2), \dots, (\lambda_n, x_n), \dots \quad (4.10)$$

which has all the properties of (2.5). Let the apriori distribution function be

$$P(\lambda_n = 1) = p, \quad P(\lambda_n = 0) = 1 - p,$$

where  $p$  is unknown.

Although the values  $x_1, x_2, \dots, x_{n-1}$  are supposed to be independent of  $\lambda_n$ , these observations do contain useful information about  $p$ ; this suggests the use of apriori observations in making decision about  $\lambda_n$ . Now, consider as a decision procedure for the sequence (4.10) any sequence  $T = \{t_n\}$  of function  $t_n = t_n(x_1, x_2, \dots, x_n)$  with values 0 and 1; in using  $T$  we take action  $a_{t_n}$  on the  $n^{\text{th}}$  component and incur the loss  $L[t_n, \lambda_n]$ . In view of (4.2) and (4.4), the Bayes risk is

$$\begin{aligned} R(t_n, p) &= E\{L[t_n, \lambda_n]\} = E\{L[0, \lambda_n] - t_n b(\lambda_n)\} \quad (4.11) \\ &= pc_1 - E[t_n b(\lambda_n)], \end{aligned}$$

where  $E$  denotes expectation w.r.t. all the random variables  $x_1, x_2, \dots, x_n$  and  $\lambda_n$ . The independent random variables  $x_1, x_2, \dots, x_n, \dots$  have the common marginal density function w.r.t.  $\mu$  on  $X$  which is given by (cf. (3.7))

$$f_p(x) = pf_1(x) + (1-p)f_0(x). \quad (4.12)$$

It follows that

$$\begin{aligned} E[t_n b(\lambda_n)] &= pE[t_n b(\lambda_n) | \lambda_n = 1] + (1-p)E[t_n b(\lambda_n) | \lambda_n = 0], \\ &= pc_1 E[t_n | \lambda_n = 1] - (1-p)c_0 E[t_n | \lambda_n = 0], \\ &= pc_1 \int \underbrace{\dots}_{x} t_n f_p(x_1) \dots f_p(x_{n-1}) f_1(x_n) d_{\mu}(x_1) \\ &\quad \dots d_{\mu}(x_n) \end{aligned}$$



$$\begin{aligned}
&= c_0(1-p) \int \overbrace{\cdot \cdot \cdot}^n_x \int t_n f_p(x_1) \\
&\quad \cdot \cdot \cdot f_p(x_{n-1}) f_0(x_n) d_\mu(x_1) \cdot \cdot \cdot d_\mu(x_n), \\
&= \int \overbrace{\cdot \cdot \cdot}^n_x \int t_n f_p(x_1) \cdot \cdot \cdot f_p(x_{n-1}) [p c_1 f_1(x_n) \\
&\quad - (1-p) c_0 f_0(x_n)] d_\mu^n,
\end{aligned}$$

where

$$d_\mu^n = d_\mu(x_1) d_\mu(x_2) \cdot \cdot \cdot d_\mu(x_n).$$

We then have

$$\begin{aligned}
E[t_n b(\lambda_n)] &= \int \left[ \int \overbrace{\cdot \cdot \cdot}^{n-1}_x \int t_n f_p(x_1) \right. \\
&\quad \left. \cdot \cdot \cdot f_p(x_{n-1}) d_\mu^{n-1} \right] \psi_p(x) d_\mu(x) \quad (4.13)
\end{aligned}$$

where  $\psi_p(x)$  is defined by (4.7).

Suppose there exists a sequence of functions  $0 < p_n = p_n(x_1, x_2, \cdot \cdot \cdot, x_n) \leq 1$  such that for every fixed  $x$  and  $\epsilon > 0$ ,

$$\lim_{n \rightarrow \infty} P[|p_n - p| > \epsilon] = 0. \quad (4.14)$$

Define

$$\begin{aligned}
t_n(x_1, \cdot \cdot \cdot, x_n) &= t_{p_n}(x_n) = 1 \quad \text{if } \psi_{p_n}(x_n) \geq 0, \quad (4.15) \\
&= 0 \quad \text{otherwise,}
\end{aligned}$$

where  $\psi_p(x)$  is a continuous function of  $p$ . Then by (4.14) we have for any  $x$  that

$$\begin{aligned}
& \int_{\mathbf{x}} \left[ \cdot \cdot \cdot \int_{\mathbf{x}}^{n-1} t_n f_p(x_1) \cdot \cdot \cdot f_p(x_{n-1}) d_{\mu}^{n-1} \right] \\
& = P[\psi_{p_n}(x) \geq 0] \longrightarrow 1 \quad \text{if } \psi_p(x) \geq 0, \\
& \longrightarrow 0 \quad \text{otherwise,}
\end{aligned}$$

and  $n \rightarrow \infty$ . Hence

$$\begin{aligned}
& \left[ \int_{\mathbf{x}} \left[ \cdot \cdot \cdot \int_{\mathbf{x}}^{n-1} t_n f_p(x_1) \cdot \cdot \cdot f_p(x_{n-1}) d_{\mu}^{n-1} \right] \psi_p(x) \right. \\
& \quad \left. \longrightarrow [\psi_p(x)]^+ \right]. \tag{4.16}
\end{aligned}$$

Further, the absolute value of the left-hand side of (4.16) is no greater than  $|\psi_p(x)|$ , and by (4.7) we have

$$\int_{\mathbf{x}} |\psi_p(x)| d_{\mu}(x) \leq pc_1 + (1-p)c_0 < \infty. \tag{4.17}$$

Thus by Lebesgue's theorem of dominated convergence, it follows from (4.16) and (4.17) that

$$\begin{aligned}
& \lim_{n \rightarrow \infty} \int_{\mathbf{x}} \left[ \int_{\mathbf{x}} \left[ \cdot \cdot \cdot \int_{\mathbf{x}}^{n-1} t_n f_p(x_1) \cdot \cdot \cdot f_p(x_{n-1}) d_{\mu}^{n-1} \right] \psi_p(x) d_{\mu}(x) \right. \\
& \quad \left. = \int_{\mathbf{x}} [\psi_p(x)]^+ d_{\mu}(x), \right.
\end{aligned}$$

or

$$\lim_{n \rightarrow \infty} E[t_n b(\lambda_n)] = \int_{\mathbf{x}} [\psi_p(x)]^+ d_{\mu}(x). \tag{4.18}$$

From (4.9) and (4.11) it follows that

$$\lim_{n \rightarrow \infty} R(t_n, p) = \lim_{n \rightarrow \infty} \left\{ pc_1 - E[t_n b(\lambda_n)] \right\},$$

$$\begin{aligned}
&= pc_1 - \int_{\mathbf{x}} [\Psi_p(\mathbf{x})]^+ d_{\mu}(\mathbf{x}) , \\
&= R(p) ,
\end{aligned}$$

so that  $T = \{t_n\}$  is asymptotically optimal in the sense that

$$\lim_{n \rightarrow \infty} R(t_n, p) = R(p).$$

The remaining problem is to exhibit a sequence  $\{p_n\}$  satisfying (4.14). Let  $h(\mathbf{x})$  be any unbiased estimator of  $\lambda(E[h(\mathbf{x})|\lambda] = \lambda)$  such that

$$\int_{\mathbf{x}} h(\mathbf{x}) f_{\lambda}(\mathbf{x}) d_{\mu}(\mathbf{x}) = \lambda .$$

Define

$$\tilde{p}_n(x_1, \dots, x_n) = \frac{1}{n} \sum_{i=1}^n h(x_i) . \quad (4.19)$$

Then since the  $x_i$ 's are independent and identically distributed random variables with

$$E[h(x_i)] = E[E[h(x_i)|\lambda]] = E[\lambda] = p, \quad (4.20)$$

it follows that

$$P\left[\lim_{n \rightarrow \infty} \tilde{p}_n = p\right] = 1 . \quad (4.21)$$

The  $\lim_{n \rightarrow \infty} h(\mathbf{x})/n \rightarrow 0$  for any fixed  $\mathbf{x}$ , so that (4.21) holds as well for the sequence  $\{\tilde{p}_n(x_1, \dots, x_{n-1}; \mathbf{x})\}$  as for (4.19). Moreover, any sequence  $\{\tilde{p}_n\}$  satisfying (4.14) can without loss of generality be restricted to values between 0 and 1, by defining

$$\begin{aligned}
p_n &= 0 \quad \text{if } \tilde{p}_n < 0, \\
&= \tilde{p}_n \quad \text{if } 0 \leq \tilde{p}_n \leq 1, \\
&= 1 \quad \text{if } \tilde{p}_n > 1.
\end{aligned} \tag{4.22}$$

We summarize these results as follows. In testing simple hypotheses in the sequential situation (4.10) where  $p$  is unknown, the decision function  $T = \{t_n\}$  defined by (4.15) is asymptotically optimal. Moreover,  $\{p_n\}$  may be any sequence  $0 \leq p_n \leq 1$  satisfying (4.14).

#### 4.3 Testing Composite Hypotheses: Known $G$

The decision function  $T$  derived for testing simple hypotheses can be generalized to the testing of composite hypotheses. Here the parameter  $\lambda$  may be any element of a general parameter space  $\Lambda$ ; there are still two actions  $a_0$  and  $a_1$ , the loss function being  $L[0, \lambda]$  for  $a_0$  and  $L[1, \lambda]$  for  $a_1$ . Again, the decision function  $t(x)$  still has values 0 and 1 defined on the sample space  $X$ . When  $x$  is observed we take action  $a_{t(x)}$  and thereby incur the loss  $L[t(x), \lambda]$ . Moreover,  $\lambda$  is still a random variable with apriori but unknown distribution function  $G$  such that

$$\int_{\Lambda} L[t(x), \lambda] dG(\lambda) < \infty. \quad (t(x) = 0, 1) \tag{4.23}$$

The Bayes risk of any decision function  $t$  is

$$\begin{aligned}
R(t, G) &= E\{L[t(x), \lambda]\} \\
&= E\{L[0, \lambda] - t(x)b(\lambda)\}
\end{aligned} \tag{4.24}$$

$$= \int_{\lambda} L[0, \lambda] dG(\lambda) - E[t(x)b(\lambda)],$$

where  $E$  denotes expectation w.r.t. the joint distribution of  $\lambda$  and  $x$ . Suppose  $f_{\lambda}$  is still defined to be the density function of  $x$  w.r.t. some measure  $\mu$  as it was in (4.1), then (4.24) can be written

$$\begin{aligned} R(t, G) &= \int_{\lambda} L[0, \lambda] dG(\lambda) - \int_x \int_{\lambda} t(x)b(\lambda)f_{\lambda}(x) dG(\lambda) d_{\mu}(x) \quad (4.25) \\ &= L[0, G] - \int_x t(x) \psi_G(x) d_{\mu}(x), \end{aligned}$$

where

$$L[0, G] = \int_{\lambda} L[0, \lambda] dG(\lambda),$$

and

$$\psi_G(x) = \int_{\lambda} b(\lambda)f_{\lambda}(x) dG(\lambda). \quad (4.26)$$

It follows that for any given  $G$ ,  $R(t, G)$  is minimized by setting  $t$  to be

$$\begin{aligned} t_G(x) &= 1 \quad \text{if } \psi_G(x) \geq 0, \\ &= 0 \quad \text{otherwise.} \end{aligned}$$

As it was in (4.8),  $t_G(x)$  is called a Bayes decision function corresponding to a given  $G$ . The minimum value of  $R(t, G)$  is

$$\begin{aligned} R(G) &= \min_t R(t, G) = R(t_G, G) \\ &= L[0, G] - \int_x [\psi_G(x)]^+ d_{\mu}(x) \end{aligned}$$

$R(G)$  is called the Bayes envelope function. If  $G$  is known,  $\psi_G$

is available and the minimum possible Bayes risk is attainable.

#### 4.4 Testing Composite Hypotheses: Unknown G

Suppose that we are in a repetitive situation in which we confront the same problem (with varying  $\lambda$  and  $x$ ) over and over again, then we have a sequence  $(\lambda_1, x_1), (\lambda_2, x_2), \dots, (\lambda_n, x_n), \dots$  of independent pairs of random variables. A sequential decision procedure  $T = \{t_n\}$  for the sequence consists of a sequence of 0, 1 valued functions  $t_n = t_n(x_1, x_2, \dots, x_n)$ . For the  $n^{\text{th}}$  component, the Bayes risk is

$$\begin{aligned} R(t_n, G) &= E\{L[t_n(x), \lambda_n]\} \\ &= L[0, G] - E[t_n b(\lambda_n)], \end{aligned}$$

where  $E$  denotes expectation w.r.t.  $x_1, \dots, x_n$  and  $\lambda_n$ . Let the random variables  $x_i$  have a common marginal density

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

Then (4.13) generalizes to give

$$\begin{aligned} E[t_n b(\lambda_n)] &= \int \underbrace{\dots}_{x} \int t_n f_G(x_1) \dots f_G(x_{n-1}) \psi_G(x) d_{\mu}^n \\ &= \int_x \left[ \int \dots \int_{x}^{n-1} t_n(x_1, \dots, x_{n-1}; x) f_G(x_1) \right. \\ &\quad \left. \dots f_G(x_{n-1}) d_{\mu}^{n-1} \right] \psi_G(x) d_{\mu}(x), \end{aligned} \tag{4.27}$$

where  $\psi_G$  is defined by (4.26).

Suppose that a sequence  $\{\psi_n\}$  of the form

$\psi_n(x_1, \dots, x_{n-1}; x)$  can be found such that for any fixed  $x$ ,

$$\lim_{n \rightarrow \infty} \psi_n(x_1, \dots, x_{n-1}; x) = \psi_G(x) . \quad (4.28)$$

We then define

$$\begin{aligned} t_n(x_1, \dots, x_{n-1}) &= 1 \text{ if } \psi_n(x_1, \dots, x_{n-1}; x) \geq 0, \\ &= 0 \text{ otherwise.} \end{aligned} \quad (4.29)$$

It follows that

$$\begin{aligned} & \left[ \int_{\mathbf{x}} \overbrace{\dots}^{n-1} t_n(x_1, \dots, x_{n-1}; x) f_G(x_1) \dots f_G(x_{n-1}) d\mu^{n-1} \right] \psi_G(x) \\ &= P[\psi_n(x_1, \dots, x_{n-1}; x) \geq 0] \psi_G(x) \rightarrow [\psi_G(x)]^+ . \end{aligned} \quad (4.30)$$

The left side of (4.30) is no greater in absolute value than  $|\psi_G(x)|$ , and by (4.23),

$$\begin{aligned} \int_{\mathbf{x}} |\psi_G(x)| d\mu(x) &\leq \int_{\mathbf{x}} \int_{\Lambda} |b(\lambda)| f_{\lambda}(x) dG(\lambda) d\mu(x) \\ &= \int_{\Lambda} |b(\lambda)| dG(\lambda) < \infty . \end{aligned} \quad (4.31)$$

By Lebesgue's theorem on dominated convergence and from (4.27), (4.30), and (4.31), it follows that

$$\lim_{n \rightarrow \infty} E[t_n b(\lambda_n)] = \int_{\mathbf{x}} [\psi_G(x)]^+ d\mu(x) ,$$

so that  $T = \{t_n\}$  defined by (4.29) is a.o. in the sense that

$$\lim_{n \rightarrow \infty} R(t_n, G) = R(G) .$$

It remains now to exhibit a sequence  $\{\psi_n\}$  satisfying (4.28),

i.e., converging to  $\psi_G$  in probability for every fixed  $x$  of  $x_n$ . The possibility of doing this will depend on the nature of the function  $\psi_G$  and on the class to which the unknown  $G$  is assumed to belong. We shall be content here with exhibiting such a sequence  $\{\psi_n\}$  for an example which is important in applications.

#### 4.5 Application--Normal Case

The Poisson case in Section 3 is an example for testing composite hypothesis for discrete case. As a second example consider the normal distribution with variance  $\sigma^2$  and unknown mean  $\lambda$ , so that

$$f_{\lambda}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\lambda)^2/2\sigma^2}.$$

The problem is to test the hypothesis  $H_0: \lambda \leq \lambda^*$  against  $H_1: \lambda > \lambda^*$ , where  $\lambda$  is a random variable with unknown distribution  $G$  on  $\Lambda = (-\infty, \infty)$  and  $\lambda^*$  is a given constant. Let  $a_i$  ( $i = 0, 1$ ) denote the action "accept  $H_i$ ", and suppose the loss in taking action  $a_0$  is given by

$$\begin{aligned} L[0, \lambda] &= 0 && \text{if } \lambda \leq \lambda^*, \\ &= \lambda - \lambda^* && \text{if } \lambda > \lambda^*, \end{aligned}$$

and the loss in taking action  $a_1$  is given by

$$\begin{aligned} L[1, \lambda] &= 0 && \text{if } \lambda > \lambda^*, \\ &= \lambda^* - \lambda && \text{if } \lambda \leq \lambda^*. \end{aligned}$$

Since



$$b(\lambda) = L[0, \lambda] - L[1, \lambda] = \lambda - \lambda^*,$$

it follows that

$$\Psi_G(x) = \int_{-\infty}^{\infty} (\lambda - \lambda^*) f_{\lambda}(x) dG(\lambda) . \quad (4.32)$$

The marginal p.d.f. of  $x$  is

$$f_G(x) = \int_{\Lambda} f_{\lambda}(x) dG(\lambda);$$

so it follows that

$$\begin{aligned} f'_G(x) &= \frac{\partial}{\partial x} \left[ \int_{\Lambda} f_{\lambda}(x) dG(\lambda) \right] \\ &= \int_{\Lambda} \frac{\partial}{\partial x} \left[ \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\lambda)^2/2\sigma^2} \right] dG(\lambda), \\ &= \int_{\Lambda} -\frac{x-\lambda}{\sigma^2} f_{\lambda}(x) dG(\lambda) , \end{aligned}$$

and hence

$$\sigma^2 f'_G(x) = - \int_{\Lambda} (x - \lambda) f_{\lambda}(x) dG(\lambda) .$$

From (4.32) we obtain

$$\begin{aligned} \Psi_G(x) &= \int_{\Lambda} (x - \lambda^*) f_{\lambda}(x) dG(\lambda) \\ &= (x - \lambda^*) \int_{\Lambda} f_{\lambda}(x) dG(\lambda) - \int_{\Lambda} (x - \lambda) f_{\lambda}(x) dG(\lambda) \\ &= (x - \lambda^*) f_G(x) + \sigma^2 f'_G(x) . \end{aligned} \quad (4.33)$$

This suggests setting

$$\begin{aligned}\psi_n(x_1, \dots, x_{n-1}:x) &= (x-\lambda^*)f_n(x_1, \dots, x_{n-1}:x) \\ &\quad + \sigma^2 g_n(x_1, \dots, x_{n-1}:x),\end{aligned}$$

and defining  $T = \{t_n\}$  by

$$\begin{aligned}t_n(x_1, \dots, x_{n-1}) &= 1, \text{ if } x_n + \frac{\sigma^2 g_n(x_1, \dots, x_{n-1}:x)}{f_n(x_1, \dots, x_{n-1}:x)} \geq \lambda^*, \\ &= 0, \text{ otherwise.}\end{aligned}\tag{4.34}$$

A satisfactory approximation to  $f_G$  and  $f'_G$  has been considered by Robbins (1963) and Samuel (1963); here, we shall only make a few remarks about this problem for the general case in which the random variable  $x$  is continuous. Let

$$F_n(x_1, \dots, x_{n-1}:x) = \frac{\text{number of terms } x_1, \dots, x_{n-1} \text{ which are } \leq x}{n}$$

be the empirical distribution function of  $x_1, \dots, x_n$ . Choose a sequence  $\{c_n\}$  of positive constants and define

$$f_n(x_1, \dots, x_{n-1}:x) = \frac{F_n(x_1, \dots, x_{n-1}:x+c_n) - F_n(x_1, \dots, x_{n-1}:x-c_n)}{2c_n}$$

as an approximation to  $f_G(x)$ . It follows that

$$P \lim_{n \rightarrow \infty} f_n(x_1, \dots, x_{n-1}:x) = f_G(x), \tag{4.35}$$

provided that  $c_n \rightarrow 0$ ,  $nc_n \rightarrow \infty$  as  $n \rightarrow \infty$ . Likewise, if we define

$$g_n(x_1, \dots, x_{n-1}; x) = \frac{f_n(x_1, \dots, x_{n-1}; x+c_n) - f_n(x_1, \dots, x_{n-1}; x-c_n)}{2c_n}$$

as an approximation to the derivative  $f'_G(x)$ , then

$$P \lim_{n \rightarrow \infty} g_n(x_1, \dots, x_{n-1}; x) = f'_G(x), \quad (4.36)$$

provided that  $c_n \rightarrow 0$ ,  $nc_n^3 \rightarrow 0$  as  $n \rightarrow \infty$  and that  $f'_G(x)$  exists, i.e., in some neighborhood of  $x$ ,

$$\left| \frac{\partial}{\partial x} f_\lambda(x) \right| \leq H(\lambda),$$

where

$$\int_{\Lambda} H(\lambda) dG(\lambda) < \infty. \quad (4.37)$$

It remains now to show that  $T = \{t_n\}$  defined by (4.34) is asymptotically optimal. Notice that (4.35), (4.36), and (4.33) together imply (4.28) provided that the condition (4.37) is verified. Now

$$\left| \frac{\partial}{\partial x} f_\lambda(x) \right| = \frac{|x-\lambda|}{\sigma^2} f_\lambda(x) \leq \frac{1}{\sqrt{2\pi}\sigma^3} |x-\lambda|,$$

so that (4.37) holds provided that

$$\int_{-\infty}^{\infty} |\lambda| dG(\lambda) < \infty. \quad (4.38)$$

But (4.38) implies (4.23), so that  $T = \{t_n\}$  defined by (4.34) is asymptotically optimal for any  $G$  for which (4.38) holds.

#### 4.6 Remarks

Suppose that, contrary to our Bayesian assumption, the parameter sequence  $\lambda_1, \lambda_2, \dots, \lambda_n, \dots$  is not a sequence of independent and identically distributed random variables with an unknown apriori distribution, but rather an arbitrary sequence of unknown constants with values 0 or 1. The expected loss on the  $n^{\text{th}}$  decision for any decision procedure  $T = \{t_n\}$  where  $t_n = t_n(x_1, \dots, x_n)$  will depend on the whole parameter vector  $\lambda_n = (\lambda_1, \lambda_2, \dots, \lambda_n)$  ( $\lambda_i = 0$  or  $1$ ). Samuel (1960) has proved that if  $\frac{d}{dp} R(p)$  exists for  $0 \leq p \leq 1$ , then  $T$  still has the property of approximating  $t_p$  which is unknown, and if  $\frac{d}{dp} R(p)$  does not exist everywhere, a suitably randomized version  $T'$  of  $T$  can be used instead.

The existence and computability of the sequence  $\{\psi_n\}$  satisfying (4.28) is an open question. The binomial distribution, due to Robbins (1963) and Samuel (1963), is an example for which  $\{\psi_n\}$  does not exist. Instead of trying to approximate to the function  $\psi_G$ , the approximation to  $G$  itself provides another way of finding a solution to the problem. Unfortunately, the theories developed for estimating  $G$  so far are still not well adapted to practical applications.

Again, throughout this section only the existence of a.o. procedures have been considered; the rate of convergence of  $R(t_n, G)$  to  $R(t, G)$  has been ignored. These and many other questions remain to be studied.

## 5. EXTENSION: SEQUENTIAL HYPOTHESIS TESTS FOR STATIONARY PROCESSES

In 1964, Tainiter applied the empirical Bayes procedures to problems of pulse detection in noisy environments; he concluded that the risk in using the asymptotically optimal sequence differed only slightly from the Bayes risk. He also pointed out that one of the difficulties in using the empirical Bayes procedures was that the noise was assumed to be uncorrelated. In view of this idea, he extended some of the theories about empirical Bayes procedures to  $r$ -dependent strictly stationary processes.

### 5.1 Definitions and Notation

Let the observed sequence of random variables be  $\{Y_n\}$ :  $n = 0, 1, 2, \dots$ ; let  $B_n = B(Y_0, Y_1, \dots, Y_n)$  be generated by  $(Y_0, \dots, Y_n)$  and  $C_n = C(Y_n, Y_{n+1}, \dots)$  be generated by  $(Y_n, Y_{n+1}, \dots)$  for all  $n \geq 0$ . The following definitions are due to Tainiter (1965).

DEFINITION: The sequence of random variables  $\{Y_n\} : n = 1, 2, \dots$ , is said to be  $r$ -dependent if  $B_n$  is independent of  $C_{n+r+1}$  for all  $n = 0, 1, \dots$ , e.g.,  $Y_{r+1}$  is independent of  $Y_0$ ,  $Y_{r+2}$  is independent of  $Y_1$ , etc.

We will further require that the sequence  $\{Y_n\} : n = 0, 1, \dots$  be strictly stationary in the usual sense.

DEFINITION: The stochastic process  $\{Y_n\} : n = 0, 1, \dots$  is said to be marginally strictly stationary if its marginal

distribution function  $F(y_{k_1}, \dots, y_{k_s})$  for any  $s$  subscripts satisfies  $F(y_{k_1}, \dots, y_{k_s}) = F(y_{k_1+h}, y_{k_2+h}, \dots, y_{k_s+h})$ ,  $h = 0, 1, 2, \dots$

Suppose the sequence  $\{Y_n\}$  is marginally strictly stationary and assume the parameters  $\{\lambda_n\}$  are still a sequence of independent identically distributed random variables with two possible values "0" and "1". Let  $P[\lambda_n = 0] = p$  and  $P[\lambda_n = 1] = 1 - p$  for  $n = 0, 1, 2, \dots$ . Let  $g(\lambda_{k_1}, \dots, \lambda_{k_s})$  be the joint density of the corresponding  $\lambda$ 's, and suppose the distribution of  $Y_k$  is  $P_0$  if  $\lambda_k = 0$  and  $P_1$  if  $\lambda_k = 1$ . Let  $F_{\lambda_{k_1}, \dots, \lambda_{k_s}}(y_{k_1}, \dots, y_{k_s})$  be the conditional distribution function of  $(Y_{k_1}, \dots, Y_{k_s})$  given  $\lambda_{k_1}, \dots, \lambda_{k_s}$ . Then by the definitions given above we have

$$F(y_{k_1}, \dots, y_{k_s}) = \sum_{i=1}^s \sum_{\lambda_{k_i}=0}^1 F_{\lambda_{k_1}, \dots, \lambda_{k_s}}(y_{k_1}, \dots, y_{k_s}) g(\lambda_{k_1}, \dots, \lambda_{k_s}),$$

where

$$g(\lambda_{k_1}, \dots, \lambda_{k_s}) = (1-p)^{\sum_{j=1}^s \lambda_{k_j}} p^{s - \sum_{j=1}^s \lambda_{k_j}}.$$

Two actions will be considered;  $a_0$  being the more appropriate when  $\lambda = 0$ , and  $a_1$  the more appropriate when  $\lambda = 1$ . To specify the loss functions, suppose that

$$\begin{aligned} L[0, 0] &= 0, & L[0, 1] &= c_1, \\ L[1, 0] &= c_0, & L[1, 1] &= 0, \end{aligned}$$

where  $c_0$  and  $c_1$  are given positive constants.

We are required to decide successively for  $n = 1, 2, \dots$  on action  $a_0$  or  $a_1$ , or, in other words, to seek a sequence of decision functions  $T = \{t_n\}$   $n = 0, 1, \dots$  which is optimal in the sense that it minimizes the Bayes risk. Note that  $Y_n (n \geq r+1)$  depends on  $(y_{n-1}, \dots, y_{n-r})$ , i.e., the decision about  $\lambda_n$  will depend on  $y_n$  and conditionally on  $(Y_{n-1}, \dots, Y_{n-r})$ .

## 5.2 Bayes Solution: Known p

Let  $F_{\lambda_n}(y_n|Y_r) = F_{\lambda_n}(y_n|y_{n-r}, \dots, y_{n-1})$  denote the conditional distribution function of  $Y_n$  given  $Y_{n-1}, \dots, Y_{n-r}$ ; let  $F(Y_{n,r}) = F(y_{n-r}, \dots, y_{n-1})$  denote the joint distribution function of  $(Y_{n-r}, \dots, Y_{n-1})$ , and let  $g(\lambda_n)$  denote the density of  $\lambda_n$ .

Since the loss on any decision, say  $t_n$ , depends on  $\lambda_n$  and  $y_n$  and is conditionally dependent on  $y_{n-r}, \dots, y_{n-1}$  ( $n > r$ ), the Bayes risk is given by

$$R(t_n, p) = \int_{Y_r} \sum_{\lambda_n=0}^1 \left[ \int_y L[t_n, \lambda_n] dF_{\lambda_n}(y_n|Y_r) \right] g(\lambda_n) dF(Y_{n,r}) \quad (5.1)$$

where  $Y_r$  is the Cartesian product space of the  $y$ 's.

It can be shown, somewhat as was done on (4.6), that (5.1) can be written

$$R(t_n, p) = (1-p)c_1 + \int_{Y_r} \int_Y [pc_0 t_n dF_0(y_n|Y_r) - (1-p)c_1 t_n dF_1(y_n|Y_r)] dF(Y_{n,r}) . \quad (5.2)$$

Without loss of generality we may suppose that the two distributions  $F_0(y_n|Y_r)$  and  $F_1(y_n|Y_r)$  are given in terms of their probability densities  $f_0, f_1$  with respect to some measure  $\mu$  on the sample space  $Y$ , so that

$$R(t_n, p) = (1-p)c_1 - \int_{Y_r} \int_Y [(1-p)c_1 f_1(y_n|Y_r) - pc_0 f_0(y_n|Y_r)] t_n d_\mu(y) d_\mu F(Y_{n,r}) , \quad (5.3)$$

which is the function we seek to minimize. Let

$$\psi_p(y_n) = (1-p)c_1 f_1(y_n|Y_r) - pc_0 f_0(y_n|Y_r) ,$$

then  $R(t_n, p)$  is minimized by using the decision function  $t_n(y_n)$  defined by

$$\begin{aligned} t_n(y_n) &= 1 \quad \text{if } \psi_p(y_n) \geq 0 , \\ &= 0 \quad \text{otherwise.} \end{aligned} \quad (5.4)$$

The corresponding Bayes risk is

$$R(p) = (1-p)c_1 - \int_{Y_r} \left[ \int_Y [\psi_p(y_n)]^+ d_\mu(y) \right] dF(Y_{n,r}) . \quad (5.5)$$

If  $p$  is known, the decision function defined by (5.4) is optimal and the minimized Bayes risk is attainable.



### 5.3 Empirical Bayes Solution: Unknown p

The problem is to find a sequence  $T^* = \{t_n^*\}$  of decision function that is a.o. in the sense that

$$\lim_{n \rightarrow \infty} R_n^*(p) = R(p) ,$$

for any unknown  $p$  such that  $0 \leq p \leq 1$ .

We shall show in a moment how to construct a sequence of functions  $0 \leq p_n = p_n(y_0, \dots, y_n) \leq 1$  such that for every fixed  $(r+1)$ -tuple  $y^{(1)}, y^{(2)}, \dots, y^{(r+1)}$  and  $\epsilon > 0$ ,

$$P \left[ \lim_{n \rightarrow \infty} |p_n(y_0, \dots, y_{n-r-1}, y^{(1)}, \dots, y^{(r+1)}) - p| > \epsilon \right] = 0. \quad (5.6)$$

Assuming this, let

$$\psi_{p_n}(y_n) = (1-p)c_1 f_1(y_n|Y_r) - p_n c_0 f_0(y_n|Y_r) ,$$

and define (cf. (5.4))

$$\begin{aligned} t_n^*(y_n) &= 1 \quad \text{if } \psi_{p_n}(y_n) \geq 0 , \\ &= 0 \quad \text{otherwise,} \end{aligned} \quad (5.7)$$

then the sequence  $T^* = \{t_n^*\}$  is asymptotically optimal. The proof is very complicated and will be omitted; it can be found in Tainiter (1965).

To exhibit a sequence  $\{p_n\}$  satisfying (5.6), let  $h(y)$  be any unbiased estimator of  $\lambda$  such that

$$\int_Y h(y) dF_\lambda(y) = \lambda , \quad (\lambda = 0, 1) ,$$

where  $F_\lambda(y)$  is the conditional distribution of  $Y_n = y$  given  $\lambda_n = \lambda$ . Define

$$\tilde{p}_n(y_0, \dots, y_n) = (n+1)^{-1} \sum_{i=0}^n h(y_i) . \quad (5.8)$$

Since the  $y$ 's are marginally identically distributed with  $E[h(y_i)] = p$ , it follows that as  $n \rightarrow \infty$ ,

$$\frac{[h(Y_0) + h(Y_1) + \dots + h(Y_n)]}{(n+1)} \rightarrow E[h(Y)],$$

so that

$$P[\lim_{n \rightarrow \infty} \tilde{p}_n = p] = 1 .$$

Thus for any fixed  $y^{(1)}, \dots, y^{(r+1)}$ , we have

$$n^{-1} \sum_{i=1}^{r+1} h(y^{(i)}) \rightarrow 0 \text{ as } n \rightarrow \infty, \text{ so that convergence still}$$

holds for the sequence

$$\tilde{p}_n(y_0, \dots, y_{n-r-1}, y^{(1)}, \dots, y^{(r+1)}) .$$

It follows that any sequence  $\{p_n\}$  satisfying (5.6) continues to do so if the value is restricted to be between 0 and 1 by defining

$$\begin{aligned} p_n &= 0 && \text{if } \tilde{p}_n < 0, \\ &= \tilde{p}_n && \text{if } 0 \leq \tilde{p}_n \leq 1, \\ &= 1 && \text{if } \tilde{p}_n > 1. \end{aligned}$$

To assume  $\{Y_n\}$  to be a sequence of stationary process is a modification to the empirical Bayes approach. In some instances these assumptions seem to be more realistic than those generally made.

## 6. CONCLUSION

In some sequential experiments it may be reasonable to assume that an apriori probability measure actually exists in the sense that the parameter is distributed according to a certain specified probability law. Neyman (1962) gave several examples to illustrate the fact of the existence and usefulness of "previous information". In general, such information will not be available to the experimenter, but the observable variables previously selected in the same way from the same population and under certain circumstances are available; these observations do contain some information about the parameter and may be used to approximate the optimal Bayes solution which could be obtained only if the apriori distribution of the parameter is completely specified.

The utilization of "previous information", in other words, the application of the empirical Bayes approach will offer certain advantages over any other approach which ignores the existence of such information. The gain in precision depends upon the nature of the uncontrollable unknown apriori distribution and may be very impressive.

The empirical Bayes approach was first established by

Robbins (1955) for estimation problems; since then, Samuel, Johns, Godambe, and Tainiter have contributed to the development of this approach. Johns (1957, 1961) has given a general discussion on nonparametric empirical Bayes procedures where the class of (conditional) probability distributions of  $X$  is not restricted to a particular parametric family. Godambe (1965) worked successfully in the field of empirical Bayes procedures in sampling finite populations; he also proved that the ratio-type estimator for the population total are "empirical Bayes" w.r.t. squared error as the loss function. Tainiter (1965) gave a complete analysis of empirical Bayes approach to  $r$ -dependent marginally stationary process.

Some of the solutions for particular problems are ready for practical applications. The following are some examples where the empirical Bayes approach seems profitable.

1. Medical Survey. It is quite reasonable to assume that the seriousness of a disease can be expressed in terms of a parameter  $\lambda$ , such that the higher the value of  $\lambda$  the more serious the disease. The parameter  $\lambda$  for an individual cannot be measured directly, but if the individual is subjected to  $r$ -independent tests and  $\lambda$  stands for the probability of a positive result, then  $\lambda$  can be regarded to be a random variable with some unknown distribution  $G$ . In this model, it seems reasonable to assume that there exists a value  $\lambda^*$  such that an individual having a value of  $\lambda$  greater than  $\lambda^*$  must be classified as sick. Here the empirical Bayes decision rule can be used to classify each of the patients as sick or healthy.

2. Quality Control. In order to decide whether a lot of  $N$  items 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. If  $\lambda$  denotes the proportion of defectives in the lot, then  $\lambda$  can be considered as a random variable which varies from lot to lot, and is distributed according to distribution function  $G$ . It has been shown by Samuel (1963) that for any fixed sample size, an empirical Bayes rule can be found. This rule will in the limit be as good as any optimal rule if  $G$  were known.

3. Pulse Detection. Pulse detection in noisy environments mentioned in Section 5 is also a practical application. If the apriori probability of no pulse is  $p$ , and of pulse is  $(1 - p)$  at each observation and  $p$  is unknown, we can use the procedure of Section 5, in particular the asymptotically optimal sequence  $T^* = \{t_n^*\}$  as given by (5.7) to classify each observation on either pulse or noise.

As of now, the empirical Bayes procedure has not been fully developed. Some important problems such as the selecting of the "best" empirical Bayes estimator, the estimation of the apriori distribution function  $G$ , the study of the rate of convergence from  $R_n(T, G)$  to  $R(G)$ , etc., still await satisfactory solutions.

## ACKNOWLEDGMENT

The writer wishes to express his heartfelt gratitude to his major professor, Dr. Ray A. Waller, for his suggestion of this topic and constant encouragement during its preparation. He also wishes to express his appreciation to Dr. Holly C. Fryer, Head of the Department of Statistics and Computer Science, Dr. A. M. Feyerherm, Professor of Statistics and Computer Science, and Dr. Ken Kemp, Professor of Statistics and Computer Science, for being members of the writer's advisory committee.

## REFERENCES

- Glivenko, V. I. (1933) Sulla determinazione empirica delle leggi di probabilita Giorn. Inst. Ital. Attuari 4.
- Godambe, V. P. (1965) Empirical Bayes Estimation in Sampling Finite Populations.
- Johns, M. V. (1957) Non-parametric Empirical Bayes Procedures. Ann. Math. Stat. 28.
- Johns, M. V. (1961) An Empirical Bayes Approach to Non-parametric Two-way Classification. Studies in Item Analysis and Predication (ed. by H. Solomn).
- Neyman, J. (1962) Two Breakthroughs in the Theory of Statistical Decision Making. International Statistical Institute, 30, 11-27.
- Robbins, H. (1955) An Empirical Bayes Approach to Statistics. Third Berkeley symposium on Mathematical Statistical and Probability.
- Robbins, H. (1963) The Empirical Bayes Approach to Testing Statistical Hypotheses. International Statistical Institute, 31, 195-208.
- Robbins, H. (1964) The Empirical Bayes Approach to Statistical Decision Problems. Ann. Math. Stat., 35, 1-20.
- Samuel, E. (1960) Asymptotic Solutions of the Sequential Compound Decision Problem Strong Convergence of the Losses of Certain Decision Rules for the Compound Decision Problem. Ann. Math. Stat., Vol. 4.
- Samuel, E. (1963) An Empirical Bayes Approach to the Testing of Certain Parametric Hypotheses. Ann. Math. Stat., 34, 1370-1403.
- Tainiter, M. (1965) Sequential Hypothesis Tests for r-Dependent Marginally Stationary Processes. Ann. Math. Stat.

EMPIRICAL BAYES APPROACH

by

KO-CHUN WANG

B. A., National Taiwan University, 1962

---

AN ABSTRACT OF A MASTER'S REPORT

submitted in partial fulfillment of the  
requirements for the degree

MASTER OF SCIENCE

Department of Statistics and Computer Science

KANSAS STATE UNIVERSITY  
Manhattan, Kansas

1970



In the Bayesian approach the parameter  $\lambda$  is considered a realization of a random variable  $\Lambda$ , distributed according to some distribution function  $G$  on  $\Omega$ . For a given  $G$  there will usually exist a Bayes rule with respect to the apriori distribution  $G$ . The use of Bayes rule will incur the minimum possible Bayes risk  $R(G)$ .

Though the assumption of the existence of an apriori distribution  $G$  is often reasonable, it is usually difficult in practice to assert what this distribution actually is; thus one will usually use a method other than a Bayes rule and incur a risk which exceeds  $R(G)$ .

The method of using previous observations to obtain a method which approaches the Bayes rule was first established by Robbins (1955), where it is called an "empirical Bayes approach". Johns (1957) generalizes the results of Robbins (1955) 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  $G$ . To describe this important property, Robbins (1964) coined the term "asymptotically optimal".

The essence of the empirical Bayes approach is that under certain circumstances prior observations may be used to construct empirical rules having the property that as the number of prior observations increases, the risk of the empirical Bayes rules converge to the risk of the Bayes rule for any apriori probability measure provided that certain moments exist.

The theory of empirical Bayes approach has not been fully developed; some of the solutions solved for particular problems are ready for practical applications, but some other problems still await rigorous investigation. Here, then, is a fruitful new field for theoretical studies.